The NCEP Climate Forecast System

S. Saha[*], S. Nadiga[*], C. Thiaw[*], J. Wang[*], W. Wang[**], Q. Zhang[**],
H. M. van den Dool[**], H.-L. Pan[*], S. Moorthi[*], D. Behringer[*], D. Stokes[*],
M. Peña*, S. Lord[*] ,G. White[*], W. Ebisuzaki[**], P. Peng[**], P. Xie[**]

Submitted to the J. Climate

Revised date : 5 Aug , 2005

[*]Environmental Modeling Center
National Centers for Environmental Prediction
NWS/NOAA/DOC, Washington, D. C.

[**]Climate Prediction Center
National Centers for Environmental Prediction
NWS/NOAA/DOC, Washington, D. C.

Corresponding author address :
Dr. Suranjana Saha, Environmental Modeling Center,
5200 Auth Road, Camp Springs, MD 20746.
E-mail : Suranjana.Saha@noaa.gov

Abstract

The Climate Forecast System (CFS), the fully coupled ocean-land-atmosphere dynamical seasonal prediction system that became operational at NCEP in August 2004, is described and evaluated in this paper. The CFS provides important advances in operational seasonal prediction on a number of fronts. For the first time in the history of U.S. operational seasonal prediction, a dynamical modeling system has demonstrated a level of skill in forecasting U.S. surface temperature and precipitation that is comparable to the skill of the statistical methods used by the NCEP Climate Prediction Center (CPC). This represents a significant improvement over the previous dynamical modeling system used at NCEP. Furthermore, the skill provided by the CFS spatially and temporally complements the skill provided by the statistical tools. The availability of a dynamical modeling tool with demonstrated skill should result in overall improvement in the operational seasonal forecasts produced by CPC.

The atmospheric component of the CFS is a lower resolution version of the Global Forecast System (GFS) that was the operational global weather prediction model at NCEP during 2003. The ocean component is the GFDL Modular Ocean Model version 3 (MOM3). There are several important improvements inherent in the new CFS relative to the previous dynamical forecast system. These include: (i) The atmosphere-ocean coupling spans almost all of the globe (as opposed to the tropical Pacific only); (ii) The CFS is a fully coupled modeling system with no flux correction (as opposed to the previous uncoupled 'tier-2' system, which employed multiple bias and flux corrections); (iii) A set of fully coupled retrospective forecasts covering a 24 year period (1981-2004),

with 15 forecasts per calendar month out to nine months into the future, have been produced with the CFS.

These 24 years of fully coupled retrospective forecasts are of paramount importance to the proper calibration (bias correction) of subsequent operational seasonal forecasts. They provide a meaningful a priori estimate of model skill that is critical in determining the utility of the real-time dynamical forecast in the operational framework. The retrospective dataset also provides a wealth of information for researchers to study interactive atmosphere-land-ocean processes.

Outline of the Paper

1. Introduction
2. Overview of the NCEP Climate Forecast System
3. Design of the CFS Retrospective Forecasts
4. CFS Performance Statistics
5. CFS Diagnostics
6. Summary, Conclusions and Discussion.
7. References
8. Appendix I: Anomaly Correlation, Systematic Error Correction and Cross Validation
9. Appendix II: Operational CFS Forecasts and Availability of CFS Data
10. Figure Legends
11. Figures

1. Introduction

It is generally assumed that the memory of the geophysical system that could aid in seasonal climate forecasting resides mainly in the ocean. The strong El Niño events of 1982/83 and 1997/98 appeared to provide empirical evidence that, at least in some cases, this is indeed true (Barnston et al. 1999). It is thus logical for the scientific community to develop global coupled atmosphere-ocean models to aid in seasonal forecasting.

At the National Centers for Environmental Prediction (NCEP) in Washington, D.C., coupled ocean-atmosphere models are looked upon as an extension of existing numerical weather prediction infrastructure. For this task, one obviously needs numerical models of both the atmosphere and the ocean, along with their own data assimilation systems. Global numerical prediction models for weather (and their attendant data assimilation systems) have matured since about 1980 and are the tool of choice today for day-to-day global weather forecasting out to one or two weeks. On the other hand, while numerical prediction models for the ocean coupled to an atmosphere have existed for a long time in research mode (Manabe and Bryan 1969), such models had not been tested

in real time forecasting, nor had a data assimilation system been developed for the ocean until the 1990's.

Ji et al. (1995) described the early data assimilation effort at NCEP (then NMC) for a tropical strip of the Pacific Ocean using the Modular Ocean Model, version 1(MOM1), developed at the Geophysical Fluid Dynamical Laboratory (GFDL) in Princeton, NJ. An ocean reanalysis was performed by Ji et al. (1995) and Behringer et al. (1998) for the Pacific basin (20°S-20°N) starting from July 1982 onward. This provided the ocean initial conditions for coupled forecast experiments, including retrospective forecasts.

The first coupled forecast model at NCEP in the mid-nineties consisted of an ocean model for the Pacific Ocean, coupled to a coarser resolution version of the then operational NMC Medium Range Forecast (MRF) atmospheric model at a spectral triangular truncation of 40 waves (T40) in the horizontal and 18 sigma levels (L18) in the vertical (Ji et al. 1994, 1998). In order to avoid very large biases, "anomaly flux corrections" were applied at the ocean-atmosphere interface. The final stand alone atmospheric forecasts were made in 'tier-2' mode, in which the sea surface temperature fields produced during the coupled integration were used, after more bias correction, as a prescribed time varying lower boundary condition for an ensemble of Atmospheric General Circulation Model (AGCM) runs. The 'tier-2' approach and its attendant flux correction procedure is common to this day. Since the SST outside the tropical Pacific had to be specified as well, damped persistence became a common substitute. This early set-up of the coupled model was known as MRFb9x for the atmospheric component, and CMP12/14 for the oceanic component. The atmospheric component was upgraded

both in physics and resolution to T62L28 several years later (Kanamitsu et al. 2002B). This upgraded system, known as the Seasonal Forecast Model (SFM) was operational at NCEP until August 2004.

Very few operational centers have been able to afford the development of a high resolution coupled atmosphere-ocean-land model (no flux correction) for real time seasonal prediction. The European Centre for Medium-Range Weather Forecasts (ECMWF) has been engaged in this effort along with NCEP. At ECMWF, the first coupled model (System-1) was developed around 1996 (Stockdale et al. 1998), with a second update (T95L40; System-2) in 2003 (Anderson et al. 2003). An evaluation against empirical models for all starting months during 1987-2001 can be found in Van Oldenborgh et al. (2003). In Australia, a coupled operational system (at R21L9 resolution) was run in retrospective mode over the period 1981-1995 from four initial months (Wang et al. 2002). In the United Kingdom, similar operational efforts have been reported in Gordon et al (2000) and Pope et al (2000). In Europe, a large research experiment was conducted recently, called the Development of a European Multimodel Ensemble system for seasonal to interannual prediction (DEMETER), in which seven different atmospheric models were coupled to about four ocean models, see Palmer et al (2004). Other quasi-operational models with flux correction include the model described in Kirtman (2003). At several other centers, such as the International Research Institute for Climate Prediction (IRI), the 'tier-2' system continues to be used (Barnston et al. 2003). In research mode, there are many more coupled models, see Schneider et al (2003) for a recent overview.

The purpose of this paper is to document the new NCEP Climate Forecast System (CFS), which became operational in August 2004. As part of the design of the CFS, three major improvements were made to the old operational coupled forecast system. First, the component models have been greatly modernized. The ocean model, MOM1, has been replaced by MOM3, and the atmospheric model, SFM, has been replaced by a coarse resolution version of the operational (as of 2003) NCEP Global Forecast System (GFS). Most notably, this change includes an upgrade in vertical resolution from the old SFM, from 28 to 64 sigma layers. Second, the ocean-atmosphere coupling is now nearly global (64°N-74°S), instead of only in the tropical Pacific Ocean, and flux correction is no longer applied. Thus, the CFS is a fully 'tier-1' forecast system. The coupling over the global ocean required an important upgrade in the ocean data assimilation as well (see Behringer et al. 2005). Third, an extensive set of retrospective forecasts ('hindcasts') was generated to cover a 24 year period (1981-2004), in order to obtain a history of the model. This history can be used operationally to calibrate and assess the skill of the real-time forecasts. Hindcast histories that were generated to assess the skill of all previous tier-2 seasonal forecast systems in use at NCEP were obtained by prescribing 'perfect' (observed) SST. This methodology is often assumed to provide an 'upper limit of predictability.' However, this method did not provide an accurate estimate of the skill of the tier-2 operational model, which used predicted, not 'perfect', SST. This methodology is still being practiced elsewhere to determine the 'skill' of multi model ensembles, etc. In the current CFS system, the model skill is assessed solely by the use of a tier-1 retrospective set of forecasts.

The first two improvements include several advances in physics and a much better coupled system, both in multi-decadal free runs (Wang et al. 2005) and in nine month forecasts from many initial conditions. Specifically, the ENSO simulation and the synoptic tropical activity (the Madden-Julian Oscillation, easterly waves, etc.) appear state of the art in the CFS with 64 vertical levels. The third item, retrospective forecasts, while costly in terms of computer time and resources, are especially important since they provide a robust measure of skill to the user of these forecasts.

The lay-out of the paper is as follows: In sections 2 and 3 we describe the components of the CFS and the organization of the retrospective forecasts respectively. In section 4 we discuss the CFS performance for its main application as a monthly/seasonal forecast tool. In section 5 we present some diagnostics highlighting strengths and systematic errors in the CFS. Summary and conclusions are found in Section 6.

2. Overview of the NCEP Climate Forecast System

The atmospheric component of the CFS is the NCEP atmospheric GFS model, as of February 2003 (Moorthi et al. 2001). Except for having a coarser horizontal resolution, it is the same as that used for operational weather forecasting with no tuning for climate applications. It adopts a spectral triangular truncation of 62 waves (T62) in the horizontal (equivalent to nearly a 200 Km Gaussian grid) and a finite differencing in the vertical with 64 sigma layers. The model top is at 0.2 hPa. This version of the GFS has been modified from the version of the NCEP model used for the NCEP/NCAR Reanalysis (Kalnay et al. 1996 ; Kistler et al. 2001), with upgrades in the parameterization of solar

radiation transfer (Hou, 1996 and Hou et al. 2002), boundary layer vertical diffusion (Hong and Pan 1996), cumulus convection (Hong and Pan 1998), gravity wave drag (Kim and Arakawa 1995). In addition, the cloud condensate is a prognostic quantity with a simple cloud microphysics parameterization (Zhao and Carr 1997, Sundqvist et al. 1989, Moorthi et al. 2001). The fractional cloud cover used for radiation is diagnostically determined by the predicted cloud condensate.

The oceanic component is the GFDL Modular Ocean Model V.3 (MOM3) (Pacanowski and Griffies 1998), which is a finite difference version of the ocean primitive equations under the assumptions of Boussinesq and hydrostatic approximations. It uses spherical coordinates in the horizontal with a staggered Arakawa B grid and the z-coordinate in the vertical. The ocean surface boundary is computed as an explicit free surface. The domain is quasi-global extending from 74°S to 64°N. The zonal resolution is 1°. The meridional resolution is 1/3° between 10°S and 10°N, gradually increasing through the tropics until becoming fixed at 1° poleward of 30°S and 30°N. There are 40 layers in the vertical with 27 layers in the upper 400 m, and the bottom depth is around 4.5 Km. The vertical resolution is 10 m from the surface to the 240-m depth, gradually increasing to about 511 m in the bottom layer. Vertical mixing follows the non-local K-profile parameterization of Large et al. (1994). The horizontal mixing of tracers uses the isoneutral method pioneered by Gent and McWilliams (1990) (see also Griffies et al. 1998). The horizontal mixing of momentum uses the nonlinear scheme of Smagorinsky (1963).

The atmospheric and oceanic components are coupled with no flux adjustment or correction. The two components exchange daily averaged quantities, such as heat and

momentum fluxes, once a day. Because of the difference in latitudinal domain, full interaction between atmospheric and oceanic components is confined to 65°S to 50°N. Poleward of 74°S and 64°N, SSTs needed for the atmospheric model are taken from observed climatology. Between 74°S and 65°S, and between 64°N and 50°N, SSTs for the atmospheric component are the weighted average of the observed climatology and the SST from the ocean component of the CFS. The weights vary linearly with latitude, such that the SSTs at 74°S and 64°N equal observed climatology and the SSTs from 65°S and 50°N equal values from the ocean component. Sea ice extent is prescribed from the observed climatology.

The ocean initial conditions were obtained from the Global Ocean Data Assimilation System (GODAS) (Behringer et al., 2005), which was made operational at NCEP in September 2003. The ocean model used in GODAS is the same as that used in the CFS retrospective forecasts. The ocean data assimilation system uses the 3-D variational technique of Derber and Rosati (1989), modified to include vertical variations in the error covariances (Behringer et al., 1998). The ocean model in GODAS was forced with weekly fluxes of heat (Q), surface buoyancy fluxes (E-P) and wind stress vectors (τ) from NCEP Reanalysis-2 (R2: Kanamitsu et al., 2002A). The GODAS sea surface temperatures were relaxed to Reynolds SST (Reynolds et al., 2002) with a time scale of 5 days. Similarly, the sea surface salinity (SSS) was relaxed to Levitus monthly climatological SSS fields (Levitus et al., 1994), but with a time scale of 10 days. The subsurface temperature data that were assimilated were obtained from expendable bathythermographs (XBTs), the tropical atmosphere-ocean (TAO) array of moored buoys, and Argo and Argo-like floats. The subsurface salinity variability strongly

influences the density stratification in the ocean through the formation of salt-stratified barrier layers, especially in the western and central equatorial Pacific Ocean (Maes and Behringer, 2000; Ji et al., 2000). Therefore, synthetic salinity data were created by imposing a climatological temperature-salinity (T-S) relationship on the observed subsurface temperature profiles, and these synthetic salinity profiles were assimilated during the ocean model runs.

For the prediction of land surface hydrology, a two layer model described in Mahrt and Pan(1984) is used in the CFS.

3. Design of the CFS Retrospective Forecasts

The CFS includes a comprehensive set of retrospective runs that are used to calibrate and evaluate the skill of its forecasts. Each run is a full nine month integration. The retrospective period covers all 12 calendar months in the 24 years from 1981 to 2004. Runs are initiated from 15 initial conditions that span each month, amounting to a total of 4320 runs. Since each run is a nine month integration, the CFS was run for an equivalent of 3240 years! Due to limitations in computer time, only 15 days in the month were used as initial conditions. These initial conditions were carefully selected to span the evolution of both the atmosphere and ocean in a continuous fashion.

The atmospheric initial conditions were from the NCEP/DOE Atmospheric Model Intercomparison Project (AMIP) II Reanalysis (R2) data (Kanamitsu et al. 2002A), and the ocean initial conditions were from the NCEP Global Ocean Data Assimilation (GODAS) (Behringer 2005). Each month was partitioned into 3 segments. The first was centered on the pentad ocean initial condition of the 11th of the month, i.e. the 5

atmospheric initial states of the 9th, 10th, 11th, 12th and 13th of the month used the same pentad ocean initial condition of the 11th. The second set of 5 atmospheric initial states of the 19th, 20th, 21st, 22nd and 23rd of the month used the same pentad ocean initial condition of the 21st of the month. The last set of 5 atmospheric initial states include the second-to-last day of the month, the last day of the month, and the 1st, 2nd and 3rd days of the next month. This last set uses the same pentad ocean initial condition of the 1st of the next month. These 15 runs from the retrospective forecasts form the ensemble that is used by operational forecasters for calibration and skill assessment for the operational monthly seasonal forecast at NCEP. Note that no perturbations of the initial conditions are applied for making the ensemble forecast. The perturbations come automatically by taking atmospheric states one day apart, but this may not be optimal.

A hypothetical example is the official monthly/seasonal forecast made by the Climate Prediction Center (CPC) of NCEP around the 10th of February. February is then considered to be the month of forecast lead zero (not issued), March is the month of forecast lead one, and so on. Runs originating from initial conditions after the 3rd of February from the retrospective forecasts are not considered for calibration of the February forecasts. This is done because in 'operations' there is a 7-day lag in obtaining the ocean initial conditions (see Appendix II for more on the design of operational CFS forecasts). The 15 members in the ensemble thus include 9th - 13th January, 19th - 23rd January, and 30th January - 3rd of February. This method of calibration and subsequent skill assessment is used throughout this paper in order to replicate the operational procedures used at NCEP and to provide the most accurate assessment possible of the CFS skill to the forecasting community.

It is important to note that the CFS model codes were 'frozen' in June 2003. The running of the entire retrospective forecasts and operational implementation of the CFS that took nearly a year, were made with these codes. No changes or tuning for results were made to these codes during the execution of the forecasts. These forecasts truly represent the "history" of the operational CFS.

4. CFS Performance Statistics

In this section we review the performance of CFS retrospective forecasts, first in terms of skill as measured by the anomaly correlation (AC) against observations (section 4.1), and then in terms of probability forecasts (section 4.2). The 'observations' used for the verification of the CFS forecasts require further explanation. Ideally, if the model analysis is good enough to be used as the initial condition, it should also be usable for verification, as is frequently done in weather forecasting. Over the years, NWP forecasts and analyses have improved hand-in-hand. However, we may not yet have reached that level of sophistication with coupled models.

i) All SST forecasts here are verified against OISST version 2 (Reynolds et al. 2002) which is the water temperature at the surface, while GODAS 'SST' is the water temperature at 5 meter below the surface. While in the tropical Pacific, SST forecast verifications against either GODAS or OIv2 are very close, the differences in verification scores are large in mid-latitudes.

ii) For surface weather elements over the continental United States (US), we use the so-called Climate Division data (Gutman and Quayle, 1996), a highly quality-controlled dataset maintained at the US National Climatic Data Center (NCDC). For temperature

over land outside the US we use the R2 fields (Kanamitsu et al. 2002), while for precipitation we use the CMAP Xie-Arkin dataset (Xie and Arkin, 1997). The R2 temperature field may not be very good, but it is unlikely that the skill of the forecast is overestimated.

iii) Verification of 500 hPa geopotential and some major atmospheric indices, such as the North Atlantic Oscillation (NAO) and the Pacific North American (PNA) pattern is done against R2 fields.

iv) The most dubious case of verification is for soil moisture. Here we use the R2 fields, and justify this as follows: a) The R2 is forced by observed precipitation, albeit with a delay of one pentad, and should be more realistic than soil moisture produced by traditional data assimilation systems, like the first Reanalysis (Kalnay et al. 1996); and b) The R2 is consistent in model formulation with the CFS forecasts. If an independent product, such as the non-operational global leaky bucket calculations (Fan and Van den Dool 2004), were to be used for verification, one first needs to transpose one product into the other, an activity fraught with difficulty.

(a) Verification of the Ensemble Mean

We refer to the Appendix I for details about definitions, and the adjustments necessary in using the Anomaly Correlation (AC) in the context of (i) systematic error correction and (ii) cross-validation, which has been adhered to in computing the results that are presented in this paper. In this section we correct for the overall mean error by subtracting the model climatology from model forecasts. See details in Appendix I.

We focus on, in order, the prediction of SST in the Nino3.4 area (5S-5N;170W-120W) of the tropical Pacific, SST in mid-latitudes, surface air temperature, precipitation, 500 hPa geopotential and soil moisture. In all cases we verify either monthly or seasonal mean values. In most cases, we verify the bias corrected ensemble mean averaged over the 15 ensemble members. In some cases, we also compare to other methods: either a previous model or some of the statistical tools that are being used by CPC. When we quote scores for 1981-2003, this includes verifying data well into 2004 for the longer lead forecasts starting in 2003.

Fig 1 shows the skill (anomaly correlation) of the Nino3.4 SST forecasts over the period 1981-2003. Nino3.4 SST is probably the single most predictable entity. We use here, and in many graphs below, a display of forecast lead in months (on the Y-axis) versus the target or verification month (on the X-axis). This makes sense when skill is more a function of target season than of lead. Forecasts for December and January exceed 0.9 in correlation for leads out to 5 months, i.e. these forecasts were initiated during the previous summer. However, forecasts for the Northern Hemisphere summer months, most notably for July, are more difficult, with correlations as low as 0.4 at leads of 7 months. The sudden drop in skill near April is known as the spring barrier.

CPC has maintained an archive of Nino3.4 SST predictions ***in real time*** since 1996. Fig. 2 shows the overall scores as a function of forecast lead for seasonal Nino3.4 SST prediction by various methods from 1997 to the present. The correlations shown are evaluated on all cases in the period. The CFS, the only retrospective method in this display, is shown by red bars. The other methods are NCEPs previous coupled model and labeled CMP14, the Canonical Correlation Analysis (CCA) (Barnston and Ropelewski

1992), Constructed Analogue (CA) (Van den Dool 1994; van den Dool and Barnston 1994), the 'Markov' method (Xue et al. 2000) and a Consolidation (CONS) of all methods available in real time (Unger et al. 1996). The CFS results, if achieved in real time, would easily have been competitive with all the other methods and, in fact, would have been a big improvement over the CMP14. A consolidation method is supposed to be at least as good as the best tool. The main reason CONS did worse than the best single method is that CMP14 scored much lower over 1997 onward than anticipated, based on its 1982-96 evaluation. This illustrates the importance of an evaluation that will hold up on independent data.

Fig 3 is the same as Fig 1, except that the ensemble mean over 14 members is verified against the one member left out. This procedure measures potential predictability under perfect model assumptions. Although scores for Nino3.4 SST are already very high in Fig.1, there is a suggestion of large improvements still ahead, especially in summer.

Ever since Barnston et al. (1994), the standard has been the performance of Nino 3.4 SST prediction at a lead of 6 months. As an example, for initial conditions in April, Fig.4 shows, as a time series, the observations and forecasts (by CFS, CMP14 and CA) for the following November. Relative to Fig.1 we note that the ensemble mean CFS not only has an overall high correlation (0.8), but also maintains amplitude in the forecast better than the other methods .

In Figs 1-3 the variable to be verified was averaged in space to be consistent with the Nino3.4 'index'. Below we report traditional skill estimates as per anomaly correlations, without any space averaging of the physical quantity. Keep in mind that

scores for the Nino3.4 area would drop by about 0.05 to 0.1 if the space averaging is not applied.

The CFS is the first NCEP coupled model for the near global oceans. While the main skill continues to be in the tropical Pacific, we can now begin to evaluate where we stand, for instance, in the prediction of mid-latitude SSTs. Fig. 5 shows at best minimal skill in SST forecasts for all target months at short leads in the Northern Hemisphere (NH), an area defined as all ocean grid points north of 35°N (no spatial mean). At all leads, the skill in the NH pales in comparison to Nino3.4, part of the problem being that GODAS and OIv2 have large differences in this area, i.e. the CFS has a very poor initial condition to begin with in mid-latitudes when OIv2 is used for verification. Against GODAS analyses, the CFS has substantial skill in mid-latitudes (not shown). See also discussion later concerning Fig.19. Fig 5 bottom shows the potential predictability of mid-latitude SST. Improvement is seen (and can be expected under perfect model assumptions) for many leads, but fundamentally the mid-latitudes appear less predictable than the tropical Pacific.

How does CFS compare to models run outside NCEP ? Restricting ourselves to 1981-2001 and using only four initial conditions in February, May, August and November, we can compare CFS to the seven European DEMETER models (Palmer et al. 2004), to which we further add CPC's constructed analogue (Van den Dool 1994). Fig. 6 (left panels) shows the anomaly correlation for the nine models in forecasting Nino3.4 SST from initial conditions in early February at lead 1 (for March) out to lead 5 (July). We have shown the CFS and CA, but marked the European models as A-G. The anomalies are relative to observed 1971-2000 climatology. The upper (left) display is for

raw forecasts, while the lower (left) display is for systematic error corrected forecasts (using cross-validation, leaving 3 years out). Forecasts from February initial conditions, just before the spring barrier, are the hardest to predict, and without bias correction some methods have a real problem. Perhaps surprisingly, the systematic error correction improves the score of many of the poor performers, indicating a linearly working system. For instance, model A improves its score from 0.2 to 0.8 correlation at lead 3. The CFS and CA have very little systematic error and do not profit from this correction. The right hand panels of Fig.6 show the results of all four initial conditions combined. The later starts (May, August and October) have much higher anomaly correlation scores than February initial conditions, so the annual mean scores look very good, and are much better than February. Annual mean scores for persistence are considerably lower than for any of the methods (after systematic error correction), but for initial conditions in November, persistence is also very high.

The SST forecasts, except for some marine interests, are not in, and of themselves, of great practical importance. The importance lies in the assumption that the SST may have considerable impact on weather elements over land. We now discuss the CFS prediction of weather elements over land, as well as 500 hPa geopotential (Z500) and soil moisture as an aid in interpreting the 'practical skill' of the CFS. Fig.7 shows the skill of monthly mean (ensemble mean) surface air temperature at 2 meters above ground (T2m) and precipitation rate (P) over the extratropical NH land north of 22.5N. Already at lead 1, these skills are extremely low, and only in summer for T2m, and winter for precipitation, is there a suggestion of non-zero correlation (for all NH grid points combined). These numbers obviously improve a little when 3-month means are used or

specific regions are considered (for US alone, see Fig 10-15 below). The reader may wonder about significance of such thoroughly low correlations. The uncertainty (sampling error) in a correlation (if small) is $1/\sqrt{(N-2)}$, where N is the effective number of cases. So a 0.4 correlation is marginally significant for time-series over 23 years. However, in Fig. 7 we aggregate over large spatial domains, such that N effectively is perhaps 50 or 100 times larger, more than enough to make even 0.05 statistically significant. This does not mean the result is practically significant, just that there is a beginning.

Fig. 8 shows similar displays of skill for monthly mean 500 hPa geopotential and soil moisture in the upper 2-meter soil. Skill for 500 hPa geopotential is quite low in NH extratropics, and only worth mentioning in the NH winter months. Fig. 8 bottom shows very high skill for soil wetness. While this skill relates mainly to high persistence, it nevertheless conveys information about the initial condition to the lower atmosphere which is known several months ahead of time. As a measure of skill, numbers in Fig.8 may be unrealistically high because independent verification is difficult to obtain. It appears, nonetheless, that the non-zero summer correlation in T2m is caused by soil wetness, while the winter skill in P is consistent with the skill in circulation (Z500). The words 'coupled model' should be thought of as including the coupling to the soil also. Potential predictability estimates (not shown) confirm the idea of skill for P in winter and T2m in summer, although we cannot report anything above 0.35 (domain averaged) in correlation, thus suggesting a low predictability ceiling.

It is common to report on skill in atmospheric teleconnections patterns. Even when overall skill (for all spatial and temporal scales) is low, the projection onto

observed patterns of the NAO and PNA may show slightly better skill by virtue of the projection onto large-scale low frequency patterns. Indeed one can amplify the scores reported in Fig.8 (top panel) for winter months by filtering the 500 hPa geopotential fields and retaining only the NAO and PNA. Fig.9 shows the time series of monthly NAO and PNA for January and February, lead 1 ensemble mean forecasts, along with the observations. We find correlations of around 0.4, which is near significance. This appears consistent with results in DEMETER (Palmer et al. 2004).

Figs. 10 and 11, bottom panels, show the spatial distribution of the anomaly correlation of the ensemble mean seasonal forecasts of T2m (Fig.10) and P (Fig.11) over the continental United States for June, July and August (JJA) on the left and December, January and February (DJF) on the right respectively. These forecasts are made at one-month lead, i.e. the summer (JJA) forecasts are made from initial conditions that range from April 9 to May 3, while the winter (DJF) forecasts are made from initial conditions that range from Oct 9 to Nov 3, for all years 1981-2003. Local correlations less than 0.3 are deemed insignificant in many CPC operational procedures. In JJA, the skill is restricted to the Northwest for both P and T2m, while most of the country has no demonstrable skill. In DJF, the skill is better, and situated mainly across the south for P, and in the middle of the country for T2m. The DJF picture of skill would be consistent with ENSO composites, i.e. mainly from years like 1982/83 and 1997/98. Skill for P in Florida in DJF is exceedingly high. Fig. 10 and 11 also address the issue of ensemble size. From top to bottom, they show the usage of 5, 10, and 15 members for making the ensemble mean, respectively. Although the pattern should stabilize more for 15 members, one can observe that, leaving details aside, the 5 member ensemble (top row) has a

similar distribution of skill in space. This is a demonstration that the CFS has reasonably stable skill.

Figs.12 and 13 are similar to Figs. 10 and 11, except that we now compare the 15-member ensemble mean (left column) to one of CPC's control statistics, the CCA (Barnston 1994) in the right column. The comparison is only coarse, since CCA is available for a much longer period (1948-2002). Fig. 12 is for lead 1 seasonal T2m and Fig. 13 is for seasonal P forecasts for the four 'official' seasons of March-April-May (MAM), June-July-August (JJA), September-October-November (SON) and December-January-February (DJF). We now face these questions: Does the CFS have any skill over the US? And if so, does it (or CCA for that matter) add any skill over and above what we know already from other methods ? The latter is a subtle (difficult) issue when skill is low and especially when there are many correlated methods. In Figs.12 and 13 it is encouraging that CFS and CCA skill do not always occur at the same geographical location, i.e. they appear complementary. Even when identical skill occurs at the same spot there is still a possibility that the skillful forecasts happen in different years due to different predictor information being exploited, such that a combination of CCA and CFS may score higher. If the source of skill is the same (and it often is) it will be hard to improve upon a single tool. This topic of consolidation (Van den Dool and Rukhovets 1994 ; Peng et al. 2002) will be the subject of future studies at NCEP and elsewhere (Doblas Reyas et al 2005). On the first question (does CFS have skill?), figures like Figs 12 and 13 should aid the CPC forecasters. In areas left blank or faint yellow there is no skill in all likelihood. In areas of correlation >=0.3 there is evidence of some skill in proportion to the correlation.

The situations most relevant to society are 'extremes', and so we close this section on ensemble mean verification with a few comments about skill of the forecasts when extremes were observed (Saha 2004). We define an extreme here as anomaly larger than 2 standard deviations. We then calculate the anomaly correlation over this small sample of cases (observed extremes of either sign). Fig. 14 shows the skill as a function of lead and target month when a monthly T2m extreme was observed in one of the four quadrants of the US (defined by 95°W and 37.5°N). Fig.15 is the same, but now for P. There is some skill, with correlations numerically higher than in Fig. 7, but noisier because the sample and the area is smaller. Correlation for extremes is just an amplified version of the regular correlation. Skill in T2m extremes is mainly in spring and fall, while skill in P resides mainly in winter, the latter in rough agreement with full NH sample results in Fig.7. This analysis is not complete. A more complete study of extremes is needed, including analysis and verification of cases where the model forecasts are extreme.

(b) Probabilistic Verification

In section 4.1, the skill of the CFS was assessed for the ensemble mean forecast as a substitute for the traditional "single" forecast, which in general, has reduced rms errors compared to the errors of the individual members (Leith 1974). In this section, we describe some measures of the performance of the entire ensemble CFS SST forecasts over the Nino 3.4 region. As an introduction, Fig 16 shows forecasts by all 15 members in 1997 and 1998 from May initial conditions. Clearly, not only is the ensemble mean close to the observation (black line), but the cloud of solutions practically rules out that

the winter 97/98 (98/99) would be anything but a warm (cold) event. These are just two cases, albeit very strong. The two measures discussed below are to describe probabilistic performance in general. The first is the reliability diagram, which indicates how forecasts probabilities correspond to the observed relative frequency of the predicted event. The second measure is the Brier Skill Score (BSS), which provides a quantitative evaluation : the mean square error of the probabilistic forecasts. (See e.g., Wilks 1995, for a description of both measures) . Probabilistic forecasts can be generated from the ensemble of forecasts by computing the fraction of members predicting a particular event among all the members in the ensemble.

BSS is defined as :

$$BSS = 1 - \frac{B}{B_c}$$

where B is the Brier Score, defined as

$$B = \frac{1}{N}\sum_{i=1}^{N}(p_i - x_i)^2$$

and B_c is the Brier score of a reference, which in this case is the observed climatological distribution. Here, p_i is the forecast probability of an 'event', x_i is its occurrence (1 if it does happen and 0 if it does not), and N is the number of realizations of the forecast process.

Three 'events' are analyzed: that the SST over Niño 3.4 is in the upper, in the middle and in the lower tercile of the distribution of the observed climatology, which we refer here to the warm, neutral and cold terciles, respectively. We first removed the mean systematic error in a 'leave-two years-out' cross-validation manner: of the 23 years of data, 21 were used to compute the mean systematic error and two were held for analysis.

Because the length of the retrospective forecast dataset is relatively short, a reduction in the number of bins was necessary to accumulate a larger sample for each of the bins and produce smooth results for the reliability diagrams. Probability bins chosen are 0-25%, 25-50%, 50-75% and 75-100%. The forecast probability assigned to these new bins is the mean forecast probability. All starting months are pooled.

1) Reliability Diagrams

Reliability diagrams were plotted for each of the forecast lead times. In Fig. 17 we show diagrams for three lead months: 1, 4 and 8 to describe the general performance of the model. Histograms showing the relative frequency of use of the forecast bins, also known as 'sharpness diagrams', are given on the right. A perfectly reliable forecast system would have its probability forecasts coinciding with the observed relative frequency and, therefore, would display a 1:1 diagonal line for each event. At lead 1, the CFS exhibits outstanding reliability to predict both the cold and neutral terciles. The warm terciles are good also but slightly overpredicted for low probabilities. That is, warm terciles are forecast with higher probabilities than what was observed. Overforecasting of the warm tercile occurs from lead 0 through lead 5. Forecasts at leads 4 and 8 still have good reliability particularly for both cold and warm terciles. Probability forecasts closer to 0 or 1 are the most reliable, whereas those for intermediate categories are more variable (in part because the sample is small). The inset histograms show that the lowest and the highest probability values are used much more often for the cold and warm terciles for all lead times. Although this is a virtue of the model, reflecting a high sharpness, there seems to be an overconfidence tendency in probabilities close to 1. This is a common problem when the ensemble spread is too narrow.

In general, the CFS ensemble has good reliability even at long lead times. Such reliability can be further improved through calibration procedures as has been done in other ensemble prediction systems (e.g., Zhu et al., 1996, Hamill and Collucci 1998, Raftery et al. 2003, Doblas-Reyes et al. 2005).

2) The Brier Skill Score

Fig. 18 shows the BSS for the prediction of three events mentioned above. The Brier score can be separated into three components: B = UNC + REL –RES, representing uncertainty, reliability and resolution, respectively (Murphy 1973, Wilks 1995). Here, UNC (= X(1-X), where X is the sample climatology of the observations) depends only on the variability of the observations and REL should be small for well-calibrated forecasts. RES should be large for a forecast system that can differentiate between situations that will or will not lead to the event in question. Fig.18 (red lines) shows high skill scores to predict (warm) upper tercile events for leads of up to 5 months, then decreasing rapidly, but still far better than climatology, for longer leads. Degraded skill after lead month 5 is mostly caused by a drop in the resolution. The skill for the middle tercile (green lines) shows a high score for the first 1-2 months of lead-time. Positive BSS are seen even at lead-time 5, but in general, the skill is much lower than the upper and lower terciles, similar to what occurs in other forecast systems (van den Dool and Toth, 1991). CFS probabilistic forecasts of the (cold) lower tercile (blue lines) has as high a skill score as the prediction of the warm tercile. Comparing these two we see that the skill for cold tercile drops earlier, which coincides with its lower resolution, than the warm for the 1 to 5 months lead. The REL term being small indicates very high reliability for all terciles.

These results are useful diagnostics for model ensemble performance. The ability of the CFS to produce probabilistic forecasts of Niño 3.4 with high skill scores for a few seasons in advance is evident.

5. CFS Diagnostics

In this section we present assorted analyses of model behavior and errors. The emphasis here is on physical interpretation and a route to possible model improvements.

(a) Model Climate Drift

The model climate drift refers to the evolution with forecast lead time of the deviation of model climatology from observations. Here the climatology for a specific season is defined as the average of the seasonal means over the retrospective period (1982-2004). For the model, the seasonal means are from the retrospective forecasts for that season. For observations, the SST is from the optimally interpolated (OISST version2) dataset (Reynolds et al. 2002), the precipitation is from the CMAP Xie-Arkin dataset (Xie and Arkin, 1997), and the 200hPa geopotential is from the Reanalysis-2 (Kanamitsu et al. 2002).

Fig. 19 exhibits the model climate drift in SST for DJF and JJA seasons and for 0-month lead, 3-month lead and 6-month lead retrospective forecasts respectively. It is evident that the bias for the DJF season in 0-month lead and 3-month lead (see Fig. 19b and c) is quite modest. In most areas of the global oceans it is less than 0.5^{o}C. Stronger bias occurs only in the small areas of the eastern equatorial Pacific and equatorial Atlantic and along the coasts (particularly the west coasts) of major continents in middle and higher latitudes. For the 6-month lead retrospective forecasts (see Fig. 19d), the bias

gets slightly stronger in the tropical Pacific and Indian oceans, indicating the tropical oceans drift away more as lead time increases. For the JJA season, the SST bias in the tropics is comparable to the DJF season, but in middle and higher latitudes, particularly in the northern hemisphere, it is much stronger and exists already at lead 0 (suggesting a major difference between GODAS and OIv2 right at the start). Warm biases with magnitudes reaching or exceeding 2^{o}C are seen across the North Pacific and North Atlantic in higher latitudes. As lead time increases from 0-month to 6-month (see Fig. 19f-h), the weak cold biases in the middle latitudes of Pacific get stronger, but interestingly, the cold biases in the equatorial Pacific becomes weaker. The warm biases along the west coasts may come from the poor parameterization of the stratiform cloud near the cold tongue regions and subtropical highs, a problem in many CGCMs.

Fig. 20 shows the same model climate drift but now for precipitation rate. Evidently the major biases are in the tropics, no matter what season or what forecast lead. For the DJF season (Fig. 20b-d), the biases are characterized by dryness in the equatorial oceans and in the South Pacific Convergence Zone (SPCZ) area, and wetness along the flanks of the dry areas. The biases get stronger as the lead time increases, similar to the situation with SST. For the JJA season (Fig. 20f-h), the dryness happens mainly in the western tropical Pacific Ocean and in the eastern Indian Ocean. The wetness patterns are similar to the DJF season, except in the South Atlantic Ocean where the errors in JJA are less.

Fig. 21 is for the model climate drift in 200 hPa eddy geopotential. As expected, the major features of the bias are in the extra tropics. For the DJF season (Fig. 21b-d),

the positive bias in the western and northwestern Pacific and its downstream wave train-like patterns suggests the geopotential bias in the northern hemisphere is tropically forced (Peng et al 2004), which is a stationary wave mode linked to tropical diabatic heating derived from the precipitation bias. For the JJA season (Fig. 21f-h), the wave train features are still discernable, though less obvious. In the Atlantic sector, there is large underestimation of the standing wave pattern.

(b) SST Bias

SST climatology has been a concern in climate simulation and prediction because latent heat anomalies, a major driving force of seasonal atmospheric circulation anomalies, depend not only on the SST anomalies but also on the time-mean condition of the ocean surface. Here we focus on the forecast bias of SST in the tropical Pacific Ocean, the most important and predictable factor affecting extratropical seasonal climate. Fig. 22 shows the 2°S-2°N average of SST bias in just the Pacific Ocean for January, April, July and October 19-23 initial conditions over the retrospective period of 23 years (1981-2003). Results from other initial days of each month are similar. There exists a large cold bias from 150°E to 110°W in target months from July to January (mainly August to October) in the forecast from January, April, and July initial conditions. A weaker cold bias is seen in target months from January to April in the forecast from July, October, and January initial conditions. Forecasts from all initial months show a warm bias close to the eastern boundary of the equatorial Pacific in target months from May to October. The causes in model physics for these SST biases are not clear. Some preliminary diagnostics indicate that the cold bias in target months from July to January in the forecast from January, April, and July initial conditions is probably associated with

the too-strong easterly momentum flux in the central eastern Pacific which results in cold temperature advection.

Fig. 23 compares forecast time-mean surface momentum flux and precipitation in June from observational analysis (panel 23a) to that of the forecast from April initial conditions (panel 23b). The differences between analysis and forecast are considered to be mainly due to model physics in the atmospheric component of the CFS because tropical SST bias in the June forecast from April initial condition is small. It is seen that the easterly momentum flux errors in the forecast appear to be associated with the ITCZ precipitation band which is too strong compared to the observational CMAP analysis. Further diagnostics and additional experiments are needed to find out which part of the model physics is responsible for the SST biases.

(c) Ocean Fields

In this section, we analyze the most important prognostic variables produced by the ocean component of the CFS: the subsurface temperature (T), zonal velocity (U), vertical velocity (W), and heat content (H). Our analysis focuses on the equatorial Pacific Ocean, the primary domain of action for the El Nino-Southern Oscillation (ENSO). Specifically, we are interested in the fidelity of reproduction of the structure of the seasonal thermocline and the zonal velocity structure of the upper ocean in the equatorial Pacific. While considerable information can be gleaned from analyses of the individual members of the 15-member ensemble, here we only plot and analyze ensemble mean fields. For this analysis, we focus on the boreal summer (JJA) and winter (DJF) climatological means in the ocean initial conditions (GODAS) and differences between the retrospective forecasts and ocean initial conditions for leads 3 and 6. A detailed

comparison of the GODAS subsurface temperature and current profiles with respect to observations is presented in Behringer and Xue (2004) for the equatorial Pacific ocean.

The climatological structure of the equatorial Pacific temperature field for 0.5°S-0.5°N is plotted for the winter (DJF) and summer (JJA) seasons for GODAS in the upper panels of Fig. 24. The plot shows the warm pool region (> 28°C) in the Western Pacific extending to 170°W in winter and to 160°W in summer. In winter, the 20°C isotherm, which is usually considered to be a proxy for the depth of the thermocline in the equatorial Pacific ocean (McPhaden et al. 1998), shoals from approximately 160 m at its deepest point in the western Pacific (140°E) to 60 m in the eastern Pacific (Nadiga et al. 2004). In the summer, the 20°C isotherm is deeper in the western Pacific than in winter, but the situation is reversed in the eastern Pacific with the shallowest depths being less than 50m at the eastern edge of the basin in summer. In winter, the SSTs are everywhere warmer than 24°C in the equatorial Pacific, but the cold tongue region extends further west than in summer. All these well-known features are well represented in the GODAS. In the middle and bottom panels in Fig.24, the differences between the retrospective forecasts and GODAS are plotted for winter (left) and summer (right) seasons for the same meridional range: 0.5°S-0.5°N. As shown in Fig. 24, the differences are small and typically less than 1°C. The greatest differences can be seen just above and below the seasonal thermocline (a dashed pink line marks the 20°C isotherm in the panels), indicating the effects of errors in vertical mixing in the ocean model. In GODAS these errors in vertical mixing are alleviated by data assimilation. Notice that the retrospective forecasts are typically colder than GODAS, except just above and below the 20°C isotherm. While the difference pattern grows in amplitude from lead 3 to lead 6 for both

winter and summer seasons, the difference pattern is quite different for winter and summer. This marked difference between winter and summer difference patterns suggests that the difference patterns do not result only from inbuilt trends/errors in the ocean model, but are also due to season-dependent errors in ocean-atmosphere coupling in the retrospective forecasts. In the eastern Pacific, these forecasts are anomalously cold compared to GODAS, and this is because of too-strong vertical upwelling in that region.

Planetary and Kelvin waves in the equatorial Pacific Ocean play an important role in setting the periodicity and duration of ENSO events (Schopf and Suarez, 1988). The zonal velocities of the planetary and Kelvin waves are functions of the depth of the thermocline locally, while the amplitudes of the mean zonal currents are functions of the slope of the thermocline. This implicit relationship between the zonal currents and computed travel times of basin-crossing planetary and Kelvin waves requires that the mean zonal velocity in the retrospective forecasts be examined for any systematic biases when compared to observations. The climatological zonal velocity in the equatorial Pacific is plotted in Fig. 25 for winter (upper-left panel) and summer (upper-right panel) seasons for GODAS for 0.5°S-0.5°N. The strong eastward velocities in the undercurrent are shown clearly in both seasons. The wind-driven westward velocities in the surface layer are stronger in winter than in summer, while the eastward velocities in the undercurrent are stronger in summer. The core of the undercurrent tilts upwards and eastwards, with the largest velocities (approximately 1m s^{-1}) being reached in the core of the undercurrent at around 100 m at 140°W. Zero zonal velocities are found in regions between the North Equatorial current (NEC) and the undercurrent and below the undercurrent. In the warm pool region, strong westward currents are found in summer,

possibly indicating the effect of Rossby waves impinging on the western edge of the Pacific Basin. In the middle and bottom panels, the differences between the retrospective forecasts and GODAS are plotted for leads 3 and 6 months for the same meridional range: 0.5°S-0.5°N . As was found in the temperature differences, we find that the difference patterns grow with lead time, and these patterns are quite different for summer and winter seasons. In winter, the largest differences are found in the undercurrent region, and the forecast eastward velocities are generally larger than GODAS. In the surface layers of the eastern Pacific, the wind-driven westward velocities are much larger than in GODAS, indicating that the surface fluxes are in error there.

The climatological vertical velocities in the equatorial Pacific Ocean are plotted in Fig. 26 for the winter (upper-left panel) and summer (upper-right panels) for GODAS for 0.5°S-0.5°N. The values shown in the figure are in mm $hour^{-1}$. The most important feature shown in the upper panels is the upwelling in the eastern Pacific. The upwelling velocities are larger in winter than in summer and reach a maximum of approximately 10 cm $hour^{-1}$. The negative vertical velocities are largest all through the water column below the warm pool, indicating the effect of downwelling Rossby waves. In both seasons, the eastward transport of mass results in strongly-positive horizontal velocity divergence in the undercurrent region. This mass transport divergence causes strong upwelling velocities above the thermocline and strong downwelling velocities below the thermocline in the eastern Pacific. The middle and bottom panels show differences between GODAS and retrospective forecasts for leads 3 and 6 months for winter and summer seasons. Unlike the difference plots for temperature and zonal velocity, the differences here are similar for winter and summer, indicating that errors in the

divergence of surface winds and not errors in oceanic mixing are the primary cause of these differences. The most striking and noticeable feature in the middle and bottom panels is the anomalously large vertical velocities in the entire water column in the eastern Pacific. In the center of the ocean basin, anomalously large negative velocities are found. The water that upwells in the eastern part of the basin is forced westward by the surface wind stress and sinks in the center of the basin. In the initial months of the forecasts, the surface layers of the eastern Pacific are colder than GODAS (see Fig. 24) as a result of anomalously large vertical upwelling. As a result of this cooling of surface layers, the negative heat advection due to vertical upwelling is reduced in the surface layers in later months, and the temperatures in the forecasts tend towards equilibrium.

The heat content of the upper ocean is an important diagnostic variable in the context of seasonal weather prediction. The recharge-discharge oscillator theory (Jin 1997) holds that anomalous buildup of heat content is a prerequisite for the occurrence of El Nino, and the equatorial Pacific ocean is purged of excess heat content during the warm event. The zonally-integrated warm water volume above the 20°C isotherm has been shown to correlate closely to the Nino 3.4 SST (McPhaden 2004). In Fig. 27, the heat content of the upper Pacific Ocean (integrated over the top 300m) is plotted for the boreal winter (upper-left panel) and summer (upper-right) seasons. The plots show the well-known double-gyre structure in the Pacific Ocean, with the maximum values recorded in the center of the gyres. The ITCZ located at approximately 10°N demarcates the northern edge of the south Pacific gyre and the southern edge of the north Pacific gyre, with approximately 5 degrees in latitude separating the two gyres. It is clear from the plot that the warm pool region contains more thermal energy in the boreal winter than

in summer. The differences between the retrospective forecasts and GODAS are plotted in the middle and bottom panels for leads 3 and 6 months, and are typically less than 5 % in most regions of the Pacific Ocean, and are especially small on the equator. The difference patterns are similar for winter and summer seasons, and the largest errors are found in the ITCZ, with the heat content in the forecasts being anomalously warm south of the ITCZ and cold north of the ITCZ. The divergence of wind stress is large in these areas, suggesting that errors in the divergence of the Ekman heat flux may be the cause of the heat content errors shown. In general, however, Fig. 27 indicates that the forecasts are able to accurately reproduce the upper ocean heat content in the equatorial Pacific ocean even for leads of 6 months and beyond. The fidelity of the forecast upper ocean heat content to the observed values is an important reason the retrospective forecast system is able to accurately predict ENSO events as is seen in Figure 28. Here, the heat content anomalies are computed and plotted for GODAS and for forecast leads of one, five and nine months. Fig. 28 clearly shows the three strongest warm ENSO events in the past 23 years (1982-83, 1987-89 and 1997-98). From the plots, it is clear that the heat content anomaly was strongest for the 1982-83 and 1997-98 ENSOs in both GODAS and forecasts. The eastward propagation of the warm anomalies by downwelling Kelvin waves during the 1982-83, 1987-88, 1991-92, 1997-98 and 2000-02 warm events is accurately reproduced in GODAS and the forecasts. Also, the 1983-84, 1988-89, 1998-99 cold events are reproduced accurately by the forecasts, and the lead 9 month forecasts reproduce all the above events, albeit with a little distortion. A careful analysis of the major events indicates a progressive lagging behind GODAS, with a maximum delay of approximately 1-2 months for the lead 9 forecasts. A noticeable exception to this

faithful reproduction of major features is the warming in the eastern Pacific in 1995, which is progressively downgraded in intensity in the retrospective forecasts, until it is no longer seen in the lead 9 forecasts.

(d) Stratosphere

Although not an item of great practical interest, in and of itself, the forecasts for the stratosphere are a great challenge scientifically because modeling the QBO is very difficult. Moreover the QBO may have an impact on the troposphere. The CFS has 25 sigma levels above 100 hPa, and it is to be hoped that this unique feature will lead to improvements in stratospheric predictions. Fig. 29 shows the skill in the CFS prediction of the QBO (defined as the stratospheric zonal mean zonal wind anomaly at the equator) as a function of lead time from zero to eight months. The verification data is taken from R2. The QBO phenomenon disappears with an e-folding of close to one year in the CFS, so one can still clearly see it at reduced amplitude in the eight month forecasts. This appears to be better than in previous models. Especially near 50 hPa, where there are many of the 64 levels in the vertical, the forecasts of the zonal mean of the zonal wind in the tropics is considerably better than persistence (which is zero after one quarter period). The 64 levels were selected to improve model prediction in the tropics, and one aspect of this improvement is the stratospheric zonal wind.

6. Summary, Conclusions and Discussion.

In this paper, we describe the new operational NCEP global coupled ocean-atmosphere model, called the Climate Forecast System or CFS. The component models are the 2003 NCEP atmospheric global weather prediction model, called the GFS, but at reduced

resolution T62L64, and the GFDL MOM3 ocean model. The coupling is once a day using daily mean fluxes. The CFS became operational in August 2004. Apart from the countless modernizations inherent in replacing the atmospheric and ocean models by newer versions, the improvements relative to the previous coupled model include specifically (i) near global atmosphere-ocean coupling (as opposed to tropical Pacific only), (ii) a fully coupled system with no flux correction (as opposed to a 'tier-2' system with multiple bias and flux corrections), and (iii) a comprehensive set of fully coupled retrospective forecasts covering the period 1981-2004, with 15 forecasts per calendar month, for forecast leads out to nine months into the future.

Since the CFS model is used for operational seasonal prediction at CPC, the 24 year retrospective forecasts, an effort which amounts to an integration of the system for nearly 3300 years, is of paramount importance for the proper calibration of subsequent real time operational forecasts.

The CFS has an acceptably low bias in tropical SST prediction, and a level of skill in forecasting Nino3.4 SST that is comparable to statistical methods used operationally at CPC, and is a large improvement over the previous operational coupled model at NCEP. Skill in predicting SST in the mid-latitudes (not done before) is much less than in the tropics, and at longer leads there is some skill only in winter. Skill for monthly and seasonal mean temperature and precipitation over NH land, and the US in particular, is modest, but still comparable to the statistical tools used operationally at CPC and not unlike a similar model at ECMWF (Van Oldenborgh et al. 2005). Skill in precipitation is mainly in winter (ENSO related), while skill in temperature is mainly in summer, when soil moisture anomalies (initialized by Reanalysis-2, which used observed precipitation

during the analysis procedure) appear helpful. Certainly the notion 'coupled' model also refers to land-atmosphere interactions.

Model behavior is reported here mainly in terms of biases or climate drift in global SST, precipitation, 200 hPa geopotential, surface wind stress, and subsurface oceanic fields. In the tropical Pacific the climate drift, while small in general, is strongest in August-September-October, even at very short lead. Some of the mid-latitude atmospheric biases appear to be forced by tropical precipitation biases. Oceanic climate drift, relative to the global ocean data assimilation, from the surface to depths of nearly 500 meters is discussed for temperature and ocean currents. In most cases, the atmospheric forcing of the ocean appears to cause climate drift in the ocean. For instance, too much upwelling in the east Pacific is caused by overly strong wind stress.

Other CFS validation efforts at NCEP, not shown but described here, show (a) much improved tropical atmosphere-ocean interactions, (b) reasonable variability around the model climate, (c) apparent skill in forecasting vertical shear in the equatorial Atlantic, and (d) improved prediction of the Quasi Biennial Oscillation (QBO). Wang et al (2005) have described the presence of an active tropical atmosphere in the CFS when using 64 layers (as opposed to 28 levels) in the vertical, and this choice of vertical levels was essential both for Nino3.4 SST simulation and low SST biases in the tropics. The variability of many fields has been studied. The overall standard deviation of monthly mean fields is reasonable for SST, T2m, and Z500, and the EOF for Z500 appear correct, at least for the first six (rotated) modes. For variability in precipitation and soil moisture, the results are not as good. Although the bias in winds over the equatorial Atlantic Ocean is considerable, the interannual variation in vertical wind shear in the Main Development

Region (MDR) for tropical hurricanes appears promising (Chelliah and Saha, 2004), and the CFS operational forecasts may aid as a new tool in the making of the operational NOAA hurricane forecast for seasonal hurricane activity for the US.

The CFS retrospective forecast data lends itself to many studies that are not just related to seasonal forecasts, and we encourage the readers to use this data. The availability of data, both the retrospective forecast data and the real time operational forecast data is discussed in Appendix II and links have been provided.

Seasonal forecasts at NCEP have been released to the public since about 1972. Initially these forecasts were made by old-fashioned subjective methods. During the 1980's and 1990's, several formal statistical tools were added to the menu that paved the way for the use of more objective methods in seasonal prediction. In these methods, an estimate of a priori skill, based on sufficient cross validation, could be used to weigh one tool versus another, before combining them into the official forecast. Adding numerical forecasts that are accompanied by appreciable a priori skill is a logical extension of this procedure. These forecasts may, or may not, have become much better, but we do have a more representative measure of a priori skill which is vital for the proper utility of seasonal forecasts.

Acknowledgements

The authors would like to recognize all the scientists and technical staff of the Global Climate and Weather Modeling Branch of EMC for their hard work and dedication to the development and implementation of the GFS. We would also like to express our thanks to the scientists at GFDL for their work in developing the MOM3 ocean model. We thank Julia Zhu, Dave Michaud, Brent Gordon and Steve Gilbert from the NCEP Central

Operations (NCO) for the timely implementation of the CFS in August 2004. George VandenBerghe and Carolyn Pasti from IBM are recognized for their critical support in the smooth running of the CFS retrospective forecasts and the operational implementation of the CFS on the NCEP IBM computers. We thank Curtis Marshall, EMC for his help in the editing of the manuscript and Åke Johansson and Augustin Vintzileos for constructive internal reviews. Finally, we thank the NOAA Office of Global Programs for the funds to obtain extra computing resources, which enabled us to complete the retrospective forecasts in a timely fashion.

References

Anderson, D. L. T., T. Stockdale, M. A. Balmaseda, L. Ferranti, F. Vitart, P. Doblas-Reyas, R. Hagedorn, T. Jung, A. Vidard, A. Troccoli, and T. Palmer, 2003: Comparison of the ECMWF seasonal forecast systems 1 and 2, including the relative performance for the 1997/8 El Nino. *Technical Memoranda 404, ECMWF*, Shinfield Park, Reading, U.K.

Barnston, A. G., 1994: Linear Statistical Short-Term Climate Predictive Skill in the Northern Hemisphere. *J. Climate*, **7**, 1513-1564

Barnston, A. G. and C. F. Ropelewski, 1992: Prediction of ENSO Episodes Using Canonical Correlation Analysis, *J. Climate*, **5**, 1316-1345.

Barnston, A.G., H. van den Dool, D. Rodenhuis, C.R. Ropelewski, V. E. Kousky, E. A. O'Lenic, R. E. Livezey, S. E. Zebiak, M. A. Cane, T. P. Barnett, N. E. Graham, Ji, Ming and A. Leetmaa, : 1994: Long-Lead Seasonal Forecasts—Where Do We Stand? *Bull. Amer. Meteor. Soc*., **75**, 2097-2114.

Barnston, A. G., A.Leetmaa, V. Kousky, R. Livezey, E.O'Lenic, H. Van den Dool, A.J. Wagner, D. Unger, 1999: NCEP Forecasts of the El Niño of 1997—98 and Its U.S. Impacts. *Bull. Amer. Meteor. Soc*., **80**, 1829-1852.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc*., **84**, 1783-1796.

Behringer, D., M. Ji, and A. Leetmaa, 1998 : An improved coupled model for ENSO prediction and implications for ocean initialization. Part I: The ocean data assimilation system. *Mon. Wea. Rev*., **126**, 1013-1021.

Behringer, D. W., et al., 2005, The Global Ocean Data Assimilation System (GODAS) at NCEP, to be submitted for publication.

Behringer, D.W., and Y. Xue, 2004: Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface, AMS 84th Annual Meeting, Seattle, Washington, 11-15.

Chelliah, M. and S. Saha, 2004 : Dynamical forecasts of atmospheric conditions associated with North Atlantic hurricane activity by the Coupled Forecast System at NCEP. *Proceedings of 29th Climate Prediction and Diagnostics Workshop, Madison, WI.*

Derber, J.D. and A. Rosati, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.,* **19**, 1333-1347.

Doblas-Reyes, F. J., R. Hagedorn and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting -- II: Calibration and combination. *Tellus A*, 57**,** 234-252

Fan, Y., and H. van den Dool, 2004 : Climate Prediction Center global monthly soil moisture data set at 0.5° resolution for 1948 to present, *J. Geophys. Res.*, **109**, D10102, doi:10.1029/2003JD004345.

Gent, P. R. and J. C. McWilliams, 1990: Isopycnal mixing in ocean circulation models. *J. Phys. Oceanogr.*, **20**, 150-155.

Gordon C., C. Cooper, C. A. Senior, H. T. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell and R. A. Wood, 2000: Simulation of SST, sea ice extents and ocean

heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics*, **16**, pp147-168

Griffies, S. M., A. Gnanadesikan, R. C. Pacanowski, V. Larichev, J. K. Dukowicz, and R. D. Smith, 1998: Isoneutral Diffusion in a z-Coordinate Ocean Model. *J. Phys. Oceanogr.*, **28**, 805-830.

Gutman, N. B. and Robert G. Quayle, 1996: A Historical Perspective of U.S. Climate Divisions. *Bulletin of the American Meteorological Society*, **77**, No. 2, 293–303.

Hamill, T. 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, *Mon. Wea. Rev.*, **129,** 550-560.

Hamill, T. M and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Hong, S.-Y. and H-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124**, 2322-2339.

Hong, S-Y and H-L. Pan, 1998: Convective Trigger Function for a Mass-Flux Cumulus Parameterization Scheme. *Mon. Wea. Rev.*, **126**, 2599–2620.

Hou, Y-T, K. A. Campana and S-K Yang, 1996: Shortwave radiation calculations in the NCEP's global model. *International Radiation Symposium*, IRS-96, August 19-24, Fairbanks, AL.

Hou, Y., S. Moorthi,. K. Campana, 2002: Parameterization of solar radiation transfer in the NCEP models. *NCEP Office Note*, **441**. http://www.emc.ncep.noaa.gov/officenotes/FullTOC.html#2000

Ji, M., A. Kumar and A. Leetmaa, 1994: A Multiseason Climate Forecast System at the National Meteorological Center. *Bull. Amer. Meteor. Soc*., **75**, Issue 4, pp.569-578

Ji, M., A. Leetmaa and J. Derber, 1995: An ocean analysis system for seasonal to interannual climate studies. *Mon. Wea. Rev*., **123**, 460-481.

Ji, M., D. W. Behringer and A. Leetmaa, 1998 : An improved coupled model for ENSO prediction and implications for ocean initialization. Part II: The coupled model. *Mon. Wea. Rev*., **126**, 1022-1034.

Ji, M., R. Reynolds and D.W. Behringer, 2000: Use of TOPEX/Poseidon sea level data for ocean analyses and ENSO prediction: some early results, *J. Climate*, 216-231.

Jin, F.-F., An equatorial recharge paradigm for ENSO., 1997: Part I: Conceptual model. *J. Atmos. Sci.* **54**, 811-829.

Kalnay, E. and Coauthors, 1996: The NCEP/NCAR 40-year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 1057-1072.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S-K. Yang, J. J. Hnilo, M. Fiorino and G. L. Potter, 2002A: NCEP–DOE AMIP-II Reanalysis (R-2) , *Bull. Amer. Meteor. Soc.*, **83**, 1631-1643.

Kanamitsu**.,** M., A. Kumar**,** J-K Schemm, H-M H. Juang, W. Wang, F. Yang, S-Y Hong, P. Peng, W. Chen and M. Ji, 2002B: NCEP dynamical seasonal forecast system 2000. *Bull. Amer. Meteor. Soc.*, **83**, 1019-1037.

Kim, Y-J and A. Arakawa, 1995: Improvement of orographic gravity wave parameterization using a mesoscale gravity wave model. *J. Atmos. Sci.*, **52**, 11, 1875-1902.

Kirtman, B. P., 2003: The COLA anomaly coupled model: Ensemble ENSO prediction. *Mon.Wea. Rev.*, **131**, 2324-2341.

Kistler, R., E. and Coauthors, 2001: The NCEP–NCAR 50–Year Reanalysis: Monthly Means CD–ROM and Documentation. *Bull. Amer. Meteor. Soc.*, **82**, No. 2, 247–268.

Large, W. G., J. C. McWilliams, and S. C. Doney, 1994: Oceanic vertical mixing: A review and a model with nonlocal boundary layer parameterization. *Rev. Geophys.*, **32**, 363-403.

Levitus, S., R. Burgett and T. P. Boyer, 1994: *Salinity*. Vol. 3, *World Ocean Atlas 1994*, NOAA Atlas NESDIS 3, U. S. Dept. of Commerce, 99pp.

Leith, C. E., 1974: Theoretical Skill of Monte Carlo Forecasts**.** *Mon. Wea. Rev.*, **102**, 409–418.

Maes, C. and D. Behringer, 2000: Using satellite-derived sea level and temperature profiles for determining the salinity variability: A new approach. *J. Geophys. Res*., **105** (C4), 8537-8547.

Mahrt, L. and H.-L. Pan, 1984: A two-layer model of soil hydrology. *Bound.-Layer Meteorol.*, **29**, 1-20.

Manabe, S. and K. Bryan, 1969: Climate Calculations with a Combined Ocean-Atmosphere Model. *J. Atmos. Sci.*, **26**, Issue 4, 786-789.

McPhaden, M.J. and Coauthors, 1998: The Tropical Ocean Global Atmospheric (TOGA) observing system: A decade of progress. *J. Geophys. Res*., **103** (C7), 14, 169-14, 240.

McPhaden, M.J., 2004: Evolution of the 2002/03 El Nino. *Bull. Amer. Meteor. Soc*., 677-695.

Moorthi, S., H.-L. Pan, P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. *NWS Technical Procedures Bulletin*, **484**, pp14. [Available at http://www.nws.noaa.gov/om/tpb/484.htm].

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Nadiga, S., J. Wang, D. Behringer and S. Saha, 2004: Ocean Retrospective Forecasts from the new Coupled Forecast System: 1981-2003. *Proceedings of 29th Climate Prediction and Diagnostics Workshop, Madison, WI.*

Palmer, T. N., and Coauthors, 2004: Development of a European Multi-Model Ensemble System for seasonal to inter-annual prediction (DEMETER). *Bull. Amer. Meteor. Soc*., **85**, 853-872.

Pacanowski, R. C. and S. M. Griffies, 1998: *MOM 3.0 Manual*, NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, USA.

P. Peng, A. Kumar, Huug van den Dool, and Anthony G. Barnston, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res*., **107** (D23), 10.1029/2002JD002712.

Peng P, Q. Zhang, A. Kumar, H. van den Dool. W. Wang, S. Saha and H-L. Pan, 2004: Variability, predictability and prediction of DJF season climate in CFS. *Proceedings of 29th Climate Prediction and Diagnostics Workshop, Madison, WI.*

Pope V.D., M. L. Gallani, P. R. Rowntree and R. A. Stratton, 2000 : The impact of new physical parametrizations in the Hadley Centre climate model - HadAM3. *Climate Dynamics*, **16**, pp123-146.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2003: Using Bayesian model averaging to calibrate forecast ensembles. *Technical Report # 440*, Department of Statistics, University of Washington. [Available online at www.stat.washington.edu/tech.reports]

Reynolds, R. W. and T. M. Smith, 1994: Improved global sea surface analyses using optimum interpolation. *J. Climate*, **7**, 929-948.

Reynolds, R.W., N. A. Rayner, T. M. Smith, D. C. Stokes and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609-1625.

Saha, S. ,2004: Validation of the NCEP CFS forecasts. *Proceedings of 29th Climate Prediction and Diagnostics Workshop, Madison, WI.*

Schneider, E. K., D. G. DeWitt., A. Rosati, B. P. Kirtman, L. Ji, and J. J. Tribbia, 2003: Retrospective ENSO forecasts: Sensitivity to atmospheric model and ocean resolution. *Mon.Wea. Rev.*, **131**, 3038-3060.

Schopf, P.S. and M. J. Suarez, 1988: Vacillations in a coupled ocean-atmosphere model. *J. Atmos. Sci*., **45**, 549-566.

Smagorinsky, J. 1963: General circulation experiments with the primitive equations: I. The basic experiment. *Mon. Wea. Rev.*, **91**, 99-164.

Stockdale, T. N., D. L. T. Anderson, J. O. S. Alves, and M. A. Balmaseda, 1998: Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature*, **392,** 370–373.

Sundqvist, H., E. Berge, and J. E. Kristjansson, 1989: Condensation and cloud studies with mesoscale numerical weather prediction model. *Mon. Wea. Rev.,* **117**, 1641-1757.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Forecast Verification: A Practitioner's Guide in Atmospheric Science, I. T. Jolliffe and D. B. Stephenson., Eds., Wiley, 137-163.

Unger, D., A. Barnston, H. Van den Dool, and V. Kousky, 1996: Consolidated forecasts of tropical Pacific SST in Niño 3.4 using two dynamical models and two statistical models. *Experimental Long-Lead Forecast Bulletin*, **5,** No. 1, 50–52.

Van den Dool, H.M., and Z. Toth, 1991: Why Do Forecasts for "Near Normal" Often Fail ? *Wea. Forecasting*, **6,** 76–85.

Van den Dool, H. M.,1994: Searching for analogues, how long must one wait ? *Tellus*, **46A**, 314-324.

Van den Dool, H. M. and A. G. Barnston, 1994: Forecasts of Global Sea Surface Temperature out to a Year using the Constructed Analogue Method. *Proceedings of 19 th Climate Diagnostics Workshop,* College Park, MD, November 14-18, 1994, 416-419.

Van den Dool, H. M. and L. Rukhovets, 1994: On the weights for an ensemble averaged 6-10 day forecast at NMC. *Weather and Forecasting*. ,**9**, 457-465.

Van Oldenborgh, G. J., M. A. Balmaseda, L. Ferranti, T. N. Stockdale and D. L. T. Anderson, 2003: Did the ECMWF seasonal forecast model outperform a statistical model over the last 15 years ? *ECMWF Technical Memorandum*, **418.**

Wang, G., R. Kleeman, N. Smith, and F. Tseitkin, 2002: The BMRC coupled general circulation model ENSO forecast system. *Mon. Wea. Rev.*, **130**, 975-991.

Wang, W., S. Saha, H.-L. Pan, S. Nadiga, and G. White, 2005: Simulation of ENSO in the new NCEP Coupled Forecast System Model. *Mon. Wea. Rev*., **133**, 1574-1593.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Xie, P., and P.A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78**, 2539 - 2558.

Xue, Y., A. Leetmaa, and M. Ji, 2000: ENSO Prediction with Markov Models: The Impact of Sea Level , *J. Climate*, **13**, 849-871.

Zhao, Q. Y., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Mon. Wea. Rev.*, **125**,1931-1953.

Zhu, Y.,G. Yyengar, Z. Toth, S. M. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, 15th Conf. on Weather Analysis and Forecasting, Norfolk, VA, Amer. Meteor. Soc., J79-J82.

Appendix I : Anomaly Correlation, Systematic Error Correction and Cross Validation.

The anomaly correlation is defined as:

$$AC = \frac{\sum\sum X'_{for}(s,t)\ X'_{obs}(s,t) / nst}{[\sum\sum X'_{for}(s,t)\ X'_{for}(s,t) / nst\ .\ \sum\sum X'_{obs}(s,t)\ X'_{obs}(s,t) / nst\ ]^{1/2}} \quad (1)$$

where, for a given lead and forecast target month/season, the summation is both over time (generally 23 cases (years)), space (e.g. the grid points north of 35^{o}N ; cosine weighting (not shown) is used in that case), and potentially even a third summation over ensemble members. nst is the number of space-time points. The primed quantities X' (X can be any variable, or ensemble mean of a variable) are defined as $X' = X - C_{obs}$ where C_{obs} is the observed climatology. In the traditional definition of AC, developed for NWP, the same observed climatology C_{obs} is removed from both forecast X_{for} and observation X_{obs}. Thus, the climatology could refer to any set of (previous 30) years (like 1971-2000) over which the climatology is traditionally calculated. In many modern studies it may seem natural, at first, to remove the model climatology C_{mdl}, if available, from X_{for} , i.e $X'_{for} = X_{for} - C_{mdl.}$ This approach can also be written $X'_{for} = X_{for} - C^*_{obs} – (C_{mdl} – C^*_{obs})$ by adding and subtracting the term C^*_{obs}, which is an observed climatology computed over the same set of years as the model climatology . The expression in parentheses is the systematic error correction, as evaluated over common years (here 1981-2003 or 1982-2004 for longer lead forecasts starting beyond May). One might say that subtracting the model climatology from X_{for} ,instead of the observed climatology, is akin to an implicit correction for the systematic error. If we use common years (e.g.1981-2003) for the observed and model climatology in equation (1) i.e. $C^*_{obs} = C_{obs}$, then the interpretation

of equation (1) is simplified and using $X'_{for} = X_{for} - C_{mdl}$ and $X'_{obs} = X_{obs} - C^*_{obs}$ in equation (1) amounts to a verification of systematic error corrected forecasts. In general, however, $C^*_{obs} \neq C_{obs}$ and thus the systematic error corrected X'_{for} in equation (1) should be kept as $X'_{for} = X_{for} - C_{obs} - (C_{mdl} - C^*_{obs})$. Furthermore, because of the implied systematic error correction, one needs to do a proper cross-validation, i.e. not use information about the year to be verified in the determination of the systematic error correction (which would amount to 'cheating'). This creates some complication in programming equation (1). The summation in time requires an 'outer loop', where each (or preferably several) years are withheld in turn. Thus, the computation of $X'_{for}(t) = X_{for}(t) - C_{obs} - (C_{mdl} - C^*_{obs})$ and $X'_{obs}(t) = X_{obs}(t) - C_{obs}$ for specific time 't' requires an adjusted C_{mdl} and C^*_{obs} (and possibly C_{obs}) such that the year 't' is not part of the various climatologies that are being computed. Cross validation (CV) is important for all verification of retrospective forecasts using anomaly correlation, rms-error and other skill measures.

It has been noted by some that the anomaly correlation is not sensitive to cross validation, which would indicate that CV is unnecessary. However this oddity of the anomaly correlation occurs only for a) CV-1-year-out (perhaps a bad practice; 3 years out is minimum) when moreover b) the observed climatology is based on the same years for which one has forecasts. When using an external climatology like 1971-2000, CV-1-year-out does change (read: lowers) the anomaly correlation.

Appendix II : Operational Forecasts and Availability of CFS Data

The initial operational implementation of CFS involved implementation of three components : Reanalysis 2 (R2) based daily atmospheric data assimilation, a daily global ocean data assimilation, and a daily nine month long coupled model integration. The retrospective forecasts with the coupled model (discussed in the preceding sections) use R2 analysis based atmospheric initial states and R2 analysis driven assimilated global ocean states. This required NCEP to make R2 atmospheric analysis operational because it is needed for both the ocean analysis and as initial condition for the CFS forecasts. This real time operational analysis is called Climate Data Assimilation System 2 (CDAS2).

The operational global ocean data assimilation system (GODAS) uses a 28-day data window symmetrically centered around the analysis time. Thus the analysis date is 14 days behind real time. So if the GODAS analysis were to be used as the initial condition, then the daily CFS forecasts would be 14 days behind real time, which would be unacceptable. Therefore, a new asymmetric GODAS which uses only 21 days of ocean data and is valid at 7 days prior to real time was developed and implemented. This asymmetric GODAS uses the previous 28-day symmetric GODAS analysis as its first guess.

Thus, operationally three analyses are performed daily in real time: the atmospheric R2 analysis is performed with analysis time which is three days before real time, the full symmetric GODAS is performed with analysis time that is 14 days before real time and an asymmetric GODAS is performed with analysis time seven days before real time. Using this asymmetric GODAS analysis as the ocean initial state and R2 analysis valid at that time as the atmospheric state (at 00 UTC), a daily CFS forecast is

made out to 9-10 months lead time. An additional daily CFS forecast with the same initial oceanic state and a slightly perturbed atmospheric state (by taking a weighted mean of the states corresponding to the real time date and a day earlier at 00 UTC) is also now operational. In all, there is at least a 60-member ensemble of CFS forecasts per month.

Monthly means of all variables and daily time-series of selected variables are being archived from these forecasts. The retrospective forecasts are used to correct the systematic bias from these monthly means before being used for the seasonal prediction.

The document at the following web link provides the necessary details on how to access the operational CFS forecast and retrospective climatological data from the official National Weather Service (NWS) data site:

http://www.emc.ncep.noaa.gov/gmb/ssaha/cfs_data/cfs_data.pdf

A web link provides the necessary details on how to access CFS retrospective time series data of monthly means of many of the most commonly used variables, as well as daily time series of selected variables from an NCEP/EMC anonymous ftp site :

http://www.emc.ncep.noaa.gov/gmb/ssaha/cfs_data/cfs_data_in_nomad.doc

Figure Legends

Fig.1 Anomaly correlation (%) of CFS ensemble mean forecasts of the monthly mean Nino3.4 SST over the period 1981-2003, as a function of target month (horizontal) and lead (vertical ; in months). Nino3.4 is defined as the spatial mean SST over 5^{o}S-5^{o}N and 170^{o}W-120^{o}W. Example, the anomaly correlation for a lead 3 forecast for March (made from 15 initial conditions beginning Nov 9^{th} and ending Dec 3^{rd}) is about 0.85. Keep in mind that the spatial averaging of SST increases the correlation relative to the traditional verification at grid points in the domain.

Fig.2 Anomaly correlation (%) by various methods of the seasonal mean Nino3.4 SST as a function of lead (horizontal; in months). The results are accumulated for all seasons in the (target) period DJF 1997/98 to DJF 2003/04. Except for CFS, all forecasts were archived in real time at CPC from 1996 onward. CMP14 is the previous coupled model, CCA is canonical correlation analysis, CA is constructed analogue, CONS is a consolidation (a weighted mean), and MARKOV is an autoregressive method (see text for references).

Fig.3 As Fig. 1 but now a 'verification' of the ensemble mean CFS (N-1 members) verified against the remaining single member. This is a predictability estimate under perfect model assumptions . Note the much reduced spring barrier scores.

Fig.4 Time series of Nino3.4 SST Anomaly (^{o}K) over the period 1981-2003. Observations for November are in the top panel and 6 month lead forecasts from April initial conditions (verifying in November) by CFS, CA and CMP14 are below. The anomaly correlation over the period is shown in the legend of each figure. Note that CFS has better amplitude than CA and CMP14. The forecasts should be considered retrospective in the years before the respective methods became operational, i.e. before 2003 for CFS, and before about 1997 for CA and CMP14.

Fig.5 Top panel: As Fig.1 but now SST grid points in Northern Hemisphere mid-latitudes (>=35^{o}N). No spatial averaging of SST is done here. Bottom panel : As Fig.3 but now potential predictability for SST in the NH (>=35^{o}N).

Fig. 6 Anomaly correlation (%) of the monthly mean Nino3.4 SST forecasts made by CFS, CA and seven European DEMETER models (A-G), as a function of lead (horizontal ; in months). Left panels are for February initial conditions. Right panels are for all initial conditions (Feb, May, Aug, Oct). Top panels are anomaly correlations of 'raw' forecasts, bottom panels are correlations for systematic error corrected forecasts.

Fig.7 Anomaly correlation (in %) of ensemble mean CFS forecasts as a function of lead and target month for monthly mean 2-meter temperature (top) and precipitation rate (bottom) over land in the NH (>=22.5^{o}N).

Fig. 8 The same as Fig.7, but now 500 hPa geopotential (top) and upper 2-meter soil moisture (bottom) in the NH. Soil moisture is over land (>=22.5^{o}N) while 500 hPa geopotential is taken north of 35^{o}N.

Fig. 9 An evaluation of skill in the CFS monthly forecast of NAO and PNA indices for January (left panels) and February (right panels) at lead 1. The forecast values (ensemble mean) are multiplied by a constant of 2.5 for the purpose of showing realistic magnitude in the anomalies compared to observations.

Fig. 10 Spatial distribution of retrospective forecast skill (anomaly correlation in %) of the ensemble mean seasonal forecasts of 2-meter temperature over the continental United States for JJA (left panels) and DJF (right panels). These forecasts are made at one-month lead, i.e. the summer (JJA) forecasts are made from initial conditions that range from April 9 to May 3, while the winter (DJF) forecasts are made from initial conditions that range from Oct 9 to Nov 3, for all years 1981-2003. From top to bottom : the number of members in the CFS ensemble mean increases from 5 to 15. Values less than 0.3 (deemed insignificant) are in faint yellow or white.

Fig. 11 As Fig.10, but now for precipitation.

Fig.12 Left panels : Spatial distribution of retrospective ensemble mean CFS forecast skill (anomaly correlation in %) for lead 1 seasonal mean 2-meter temperature over the continental United States. The target seasons are, from top to bottom, MAM, JJA, SON and DJF. Right panels are the same, but for CCA. Note that CCA is based on a longer period, 1948-2003. Correlations less than 0.3 are in faint yellow and white.

Fig. 13 The same as Fig.12, but now precipitation.

Fig. 14 Anomaly correlation (in %) of ensemble mean CFS forecasts as a function of lead (vertical) and target month (horizontal) for monthly mean 2-meter temperature over four quadrants of the continental United States (using 95^{o}W and 37.5^{o}N to define quadrants, see map at the top), evaluated only over those instances during 1981-2003 when an anomaly larger than 2 standard deviation occurred in the observations (anywhere in the quadrant). The much reduced sample size (relative to Fig.7), causes noisier patterns.

Fig. 15 The same as Fig. 14, but now precipitation.

Fig. 16 Forecast plumes of Nino3.4 SST anomalies (^{o}K) from 15 initial conditions (from April 9 to May 3) in 1997 (top) and 1998 (bottom). All members are shown by red dotted lines, the ensemble mean is a full red line, and the observations are shown as a full black line.

Fig. 17 Reliability diagrams of CFS forecast probabilities that Niño 3.4 SST predictions falls in the upper (red), the middle (green) and lower (blue) terciles of the observed climatology for (top) lead 1, (middle) lead 4 and (bottom) lead 8 months. The histograms on the right indicate the frequency of forecasts with probabilities in the ranges 0.0-0.25, 0.25-0.50, 0.50-0.75 and 0.75-1.0. Red colors correspond to forecasts for the upper (warm), green to the middle (neutral) and

blue to the lower (cold) terciles. The black line (perfect reliability) is for reference.

Fig. 18 Brier skill score (BSS), full lines, Reliability (dash-dot) and Resolution (dashed) as a function of lead time for three events: that SST in Nino3.4 is in the above tercile (red), in the middle tercile (green) and lower tercile (blue curves).

Fig.19. Observed climatology and the CFS model climate drift for SST. The climatology is defined over the period of 1982-2004. The climate drift is obtained by subtracting the observed climatology from the model forecast climatology. Left panels are for the winter season (DJF) and right panels are for the summer season (JJA). The top panels are the observed climatology. The middle and lower panels are the model climate drift for the 0, 3 and 6-month lead, respectively. Unit is ^{o}C.

Fig.20. Same as Fig.19 but for precipitation rate. Unit is mm/day.

Fig. 21. Same as Fig.19 but for 200hPa geopotential with the zonal mean removed. Unit is meters.

Fig. 22. Climate drift (Bias) of 2^{o}S-2^{o}N average SSTs in the Pacific for forecast from initial conditions of (a) January, (b) April, (c) July, and (d) October. Contours are

drawn at 0.5 K interval. Negative values are shaded. The SST bias is relative to monthly OIv2 fields averaged over 1982-2004.

Fig. 23. Precipitation rate (color shadings) and surface momentum flux (vectors) for June from (a) R2/CMAP, and (b) CFS forecast from April initial condition. Contours are the amplitude of surface momentum flux (0.1 N m^{-2}). Precipitation rate is shaded at 1, 2, 4, 8, 16, and 20 mm day^{-1}.

Fig. 24: The climatology of GODAS subsurface temperature in a depth-longitude cross section along the Equator in the Pacific and mean difference between the forecasts and GODAS in degrees Celsius. The figures on the left are for boreal winter (DJF), while the figures on the right are for boreal summer (JJA). The top panels show the climatology of subsurface temperature from GODAS. Note that a different scale is used for the color bar in the top panel. In the middle and bottom panels, the dashed pink line marks the 20C isotherm.

Fig. 25: As Fig. 24 but now zonal velocity in cm/s.

Fig. 26: As Fig. 24, but now vertical velocity in mm/hour.

Fig. 27: As Fig. 24, but now a latitude/longitude representation of the upper ocean heat content in 10^7 J m^{-2}.

Fig. 28 Longitude-time plots of heat content anomalies along the equator in the Pacific from GODAS and CFS retrospective predictions. The climatology was computed for the period: 1982-2003. Unit is 10^7 J m^{-2}.

Fig. 29: Anomaly correlation (%) of Zonal mean zonal wind anomaly at the equator as a function of pressure level (above 100 hPa) versus forecast lead time (in month).

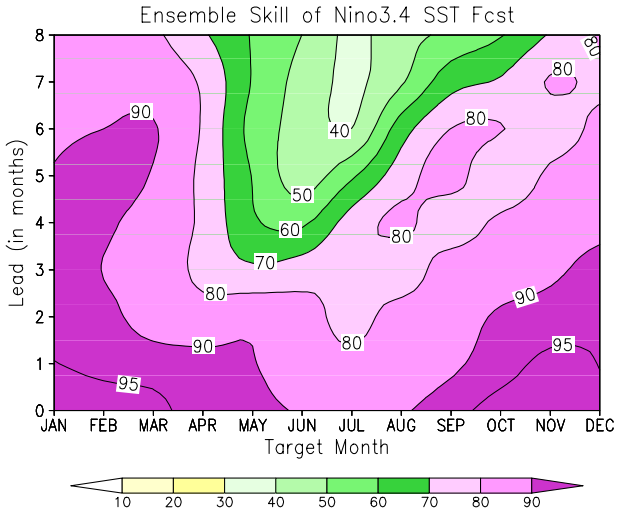

Fig.1 Anomaly correlation (%) of CFS ensemble mean forecasts of the monthly mean Nino3.4 SST over the period 1981-2003, as a function of target month (horizontal) and lead (vertical; in months). Nino3.4 is defined as the spatial mean SST over 5°S-5°N and 170°W-120°W. Example, the anomaly correlation for a lead 3 forecast for March (made from 15 initial conditions beginning Nov 9th and ending Dec 3rd) is about 0.85. Keep in mind that the spatial averaging of SST increases the correlation relative to the traditional verification at grid points in the domain.

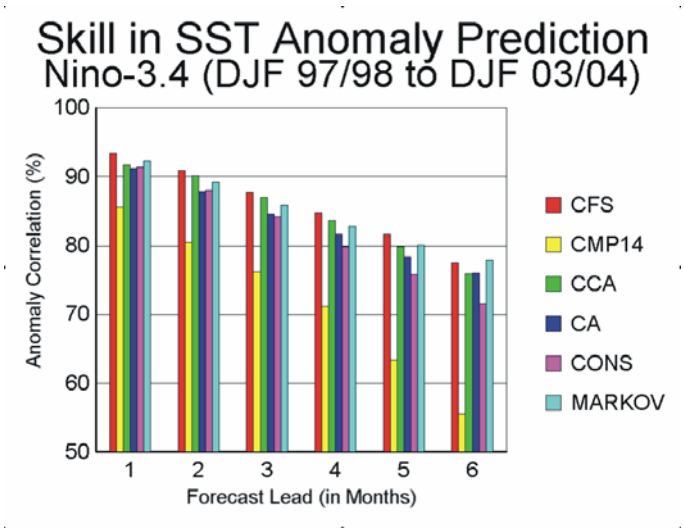

Fig.2 Anomaly correlation (%) by various methods of the seasonal mean Nino3.4 SST as a function of lead (horizontal; in months). The results are accumulated for all seasons in the (target) period DJF 1997/98 to DJF 2003/04. Except for CFS, all forecasts were archived in real time at CPC from 1996 onward. CMP14 is the previous coupled model, CCA is canonical correlation analysis, CA is constructed analogue, CONS is a consolidation (a weighted mean), and MARKOV is an autoregressive method (see text for references).

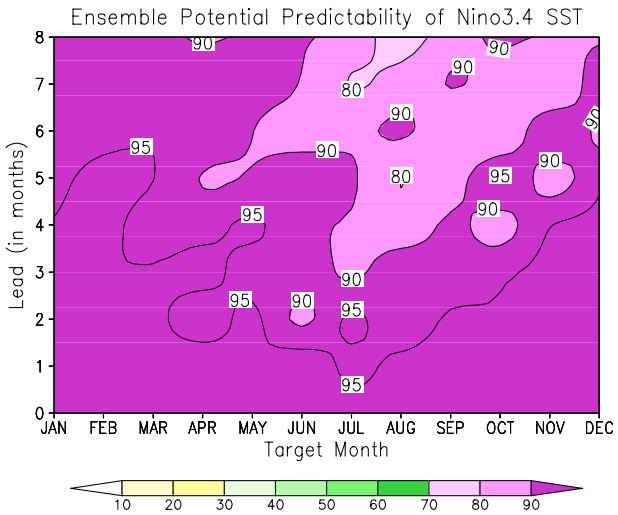

Fig.3 As Fig. 1 but now a 'verification' of the ensemble mean CFS (N-1 members) verified against the remaining single member. This is a predictability estimate under perfect model assumptions . Note the much reduced spring barrier scores.

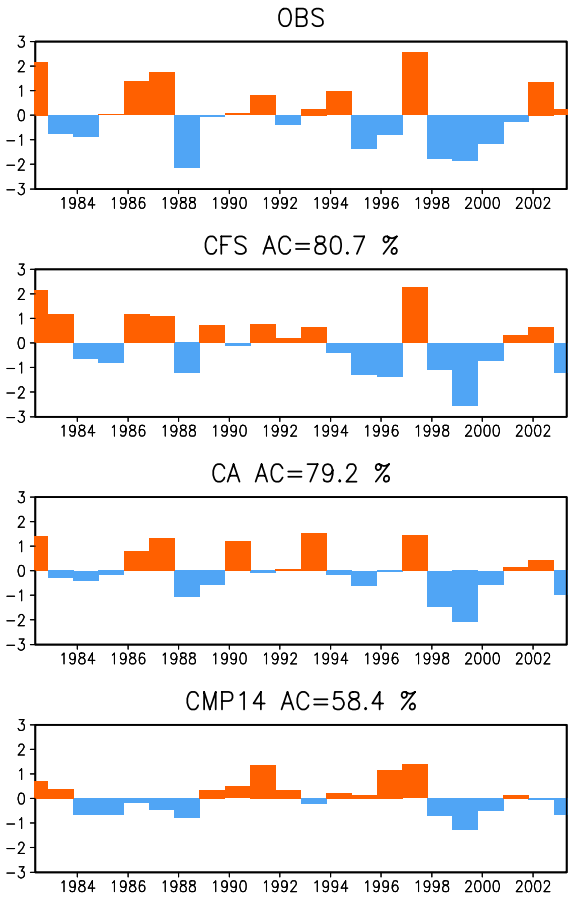

Fig.4 Time series of Nino3.4 SST Anomaly (^{o}K) over the period 1981-2003. Observations for November are in the top panel and 6 month lead forecasts from April initial conditions (verifying in November) by CFS, CA and CMP14 are below. The anomaly correlation over the period is shown in the legend of each figure. Note that CFS has better amplitude than CA and CMP14. The forecasts should be considered retrospective in the years before the respective methods became operational, i.e. before 2003 for CFS, and before about 1997 for CA and CMP14.

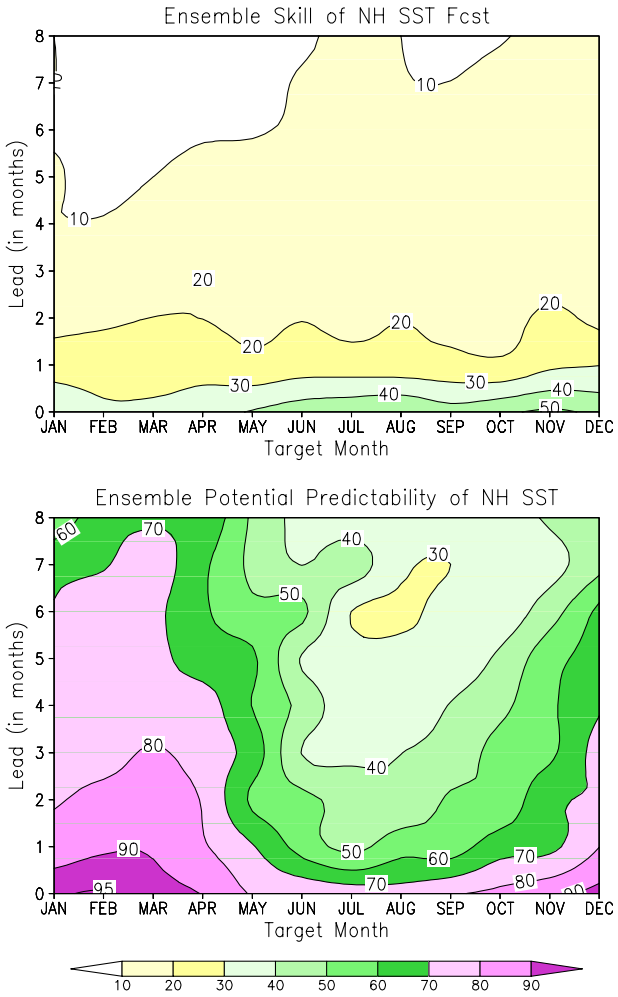

Fig.5 Top panel: As Fig.1 but now SST grid points in Northern Hemisphere mid-latitudes (>=35^{o}N). No spatial averaging of SST is done here. Bottom panel : As Fig.3 but now potential predictability for SST in the NH (>=35^{o}N).

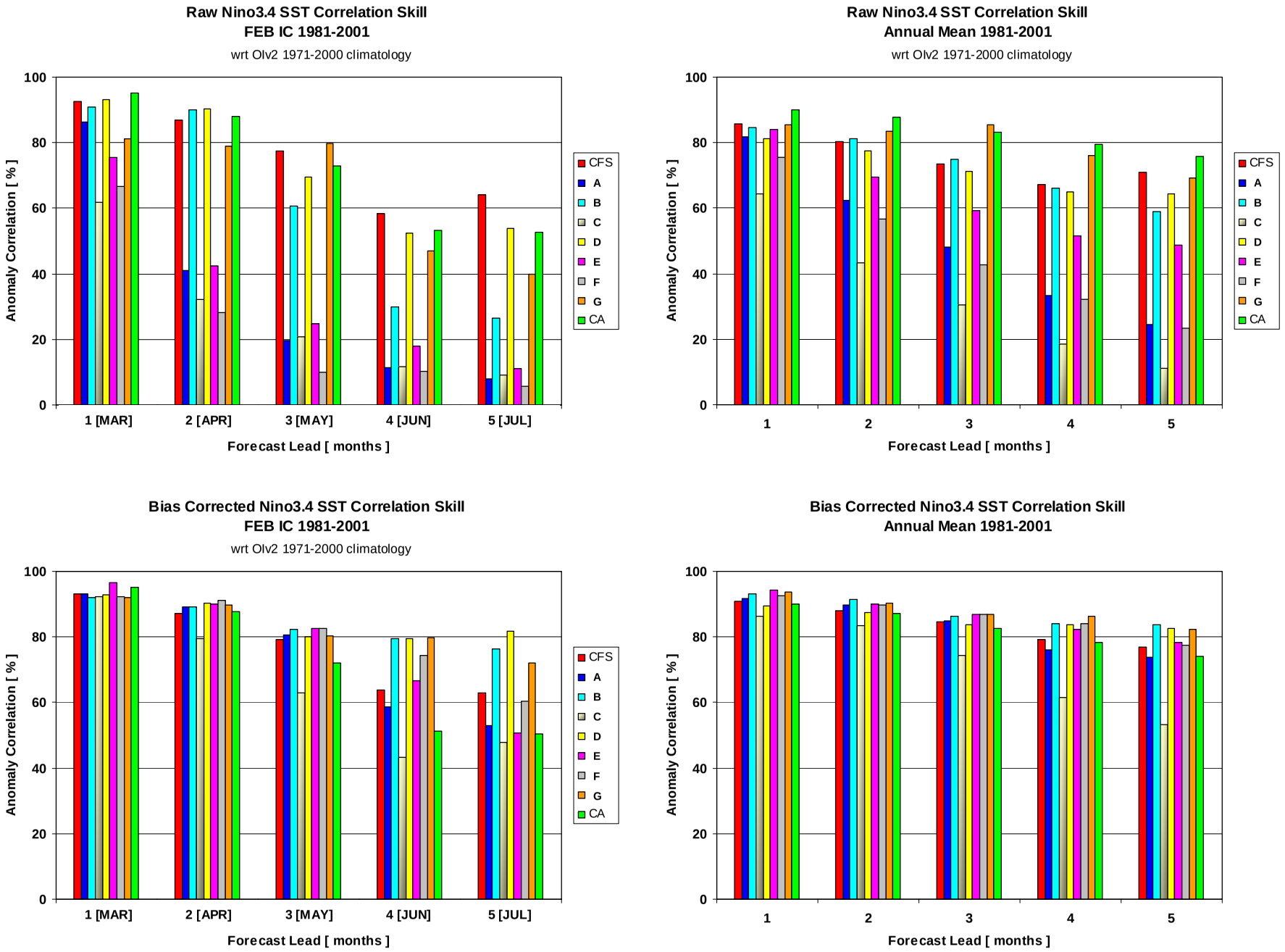

Fig. 6 Anomaly correlation (%) of the monthly mean Nino3.4 SST forecasts made by CFS, CA and seven European DEMETER models (A-G), as a function of lead (horizontal ; in months). Left panels are for February initial conditions. Right panels are for all initial conditions (Feb, May, Aug, Oct). Top panels are anomaly correlations of 'raw' forecasts, bottom panels are correlations for systematic error corrected forecasts.

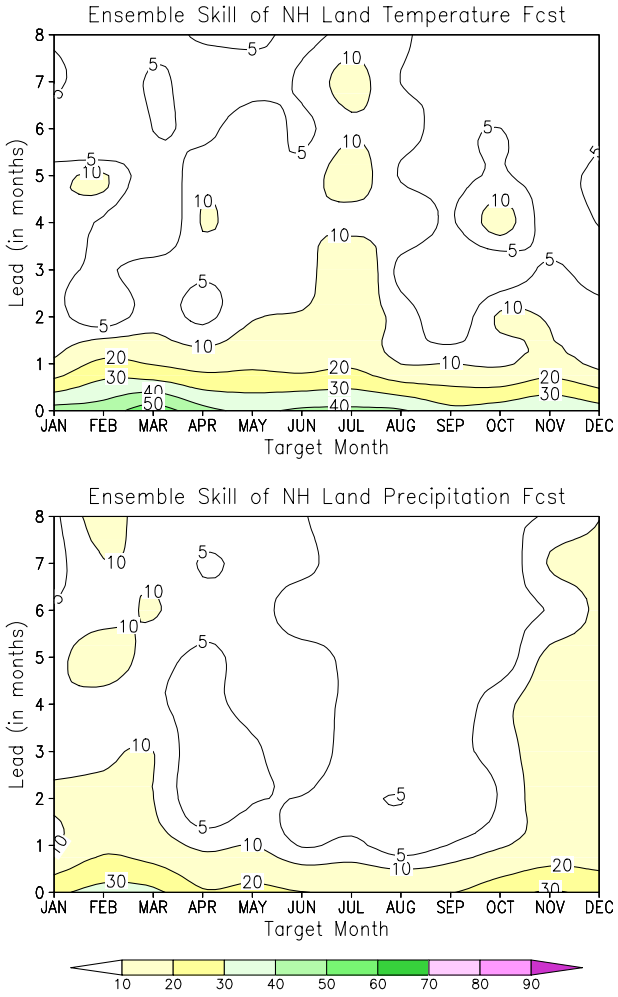

Fig.7 Anomaly correlation (in %) of ensemble mean CFS forecasts as a function of lead and target month for monthly mean 2-meter temperature (top) and precipitation rate (bottom) over land in the NH (>=22.5^{o}N).

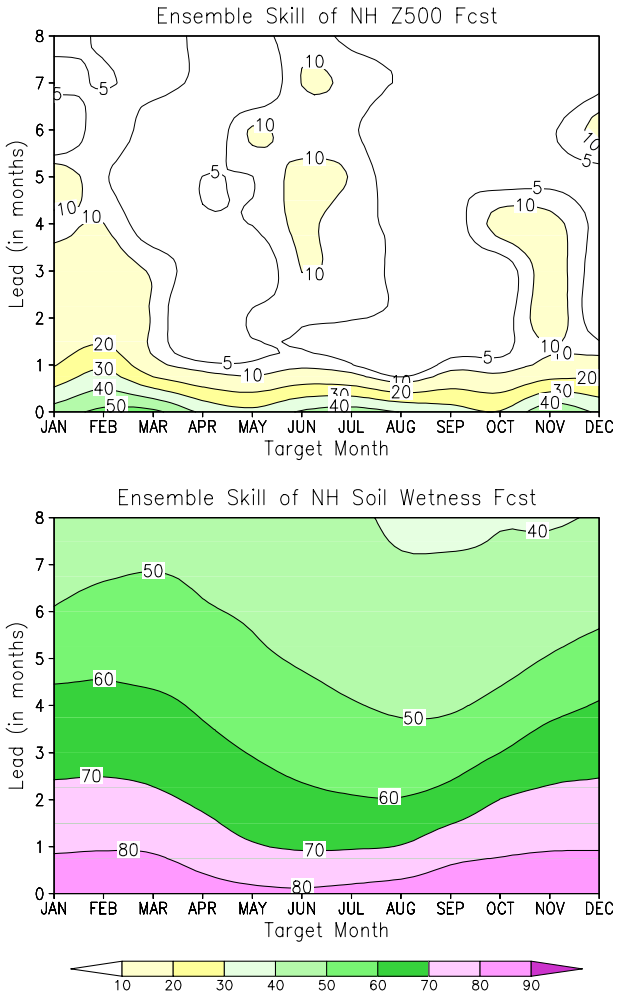

Fig. 8 The same as Fig.7, but now 500 hPa geopotential (top) and upper 2-meter soil moisture (bottom) in the NH. Soil moisture is over land (>=22.5^{o}N) while 500 hPa geopotential is taken north of 35^{o}N.

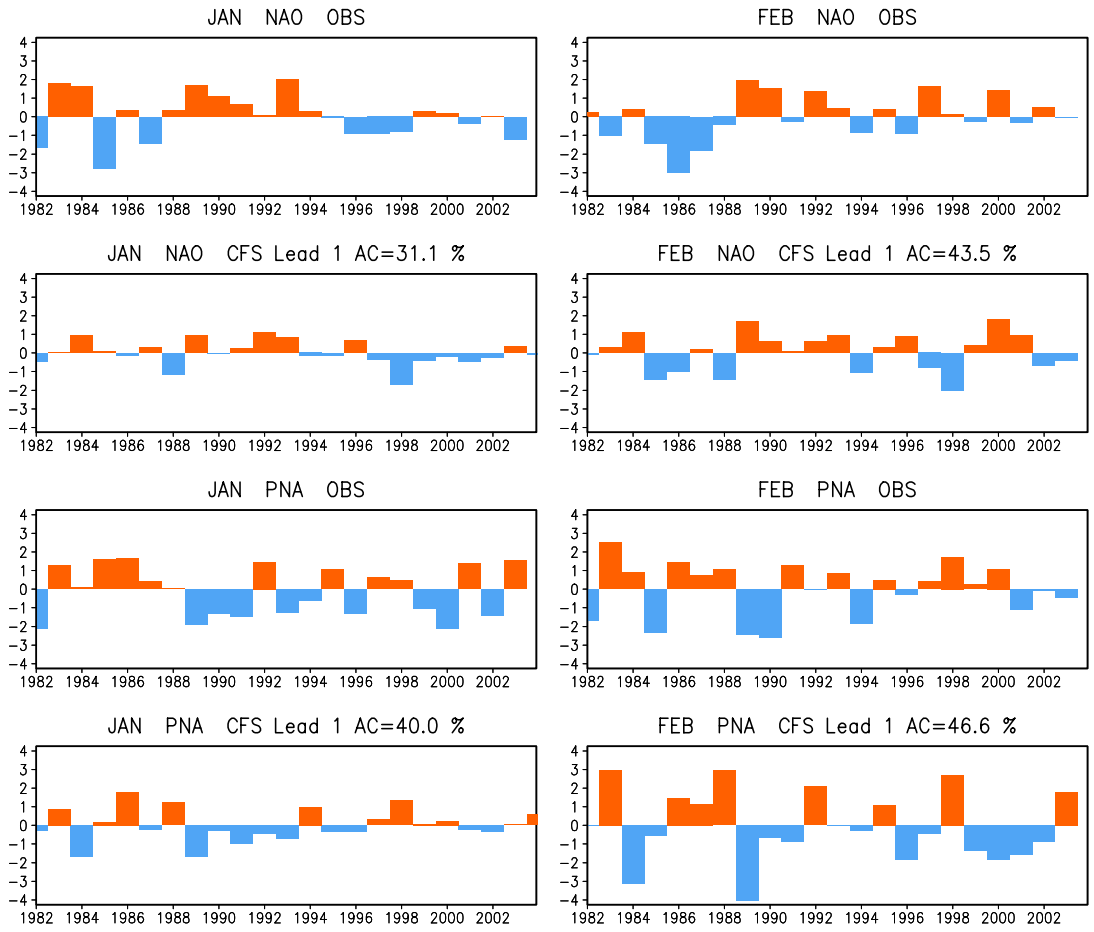

Fig. 9 An evaluation of skill in the CFS monthly forecast of NAO and PNA indices for January (left panels) and February (right panels) at lead 1. The forecast values (ensemble mean) are multiplied by a constant of 2.5 for the purpose of showing realistic magnitude in the anomalies compared to observations.

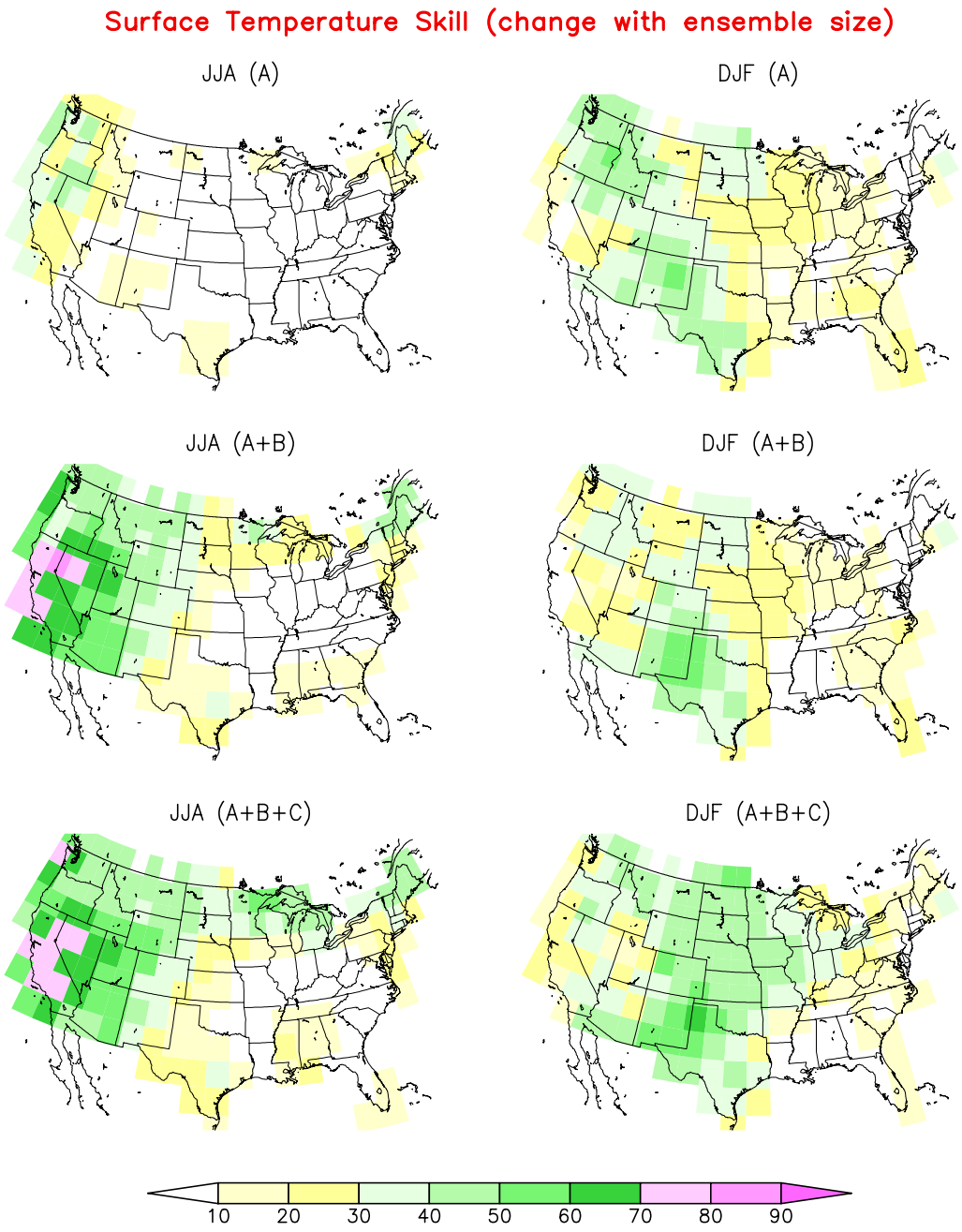

Fig. 10 Spatial distribution of retrospective forecast skill (anomaly correlation in %) of the ensemble mean seasonal forecasts of 2-meter temperature over the continental United States for JJA (left panels) and DJF (right panels). These forecasts are made at one-month lead, i.e. the summer (JJA) forecasts are made from initial conditions that range from April 9 to May 3, while the winter (DJF) forecasts are made from initial conditions that range from Oct 9 to Nov 3, for all years 1981-2003. From top to bottom : the number of members in the CFS ensemble mean increases from 5 to 15. Values less than 0.3 (deemed insignificant) are in faint yellow or white.

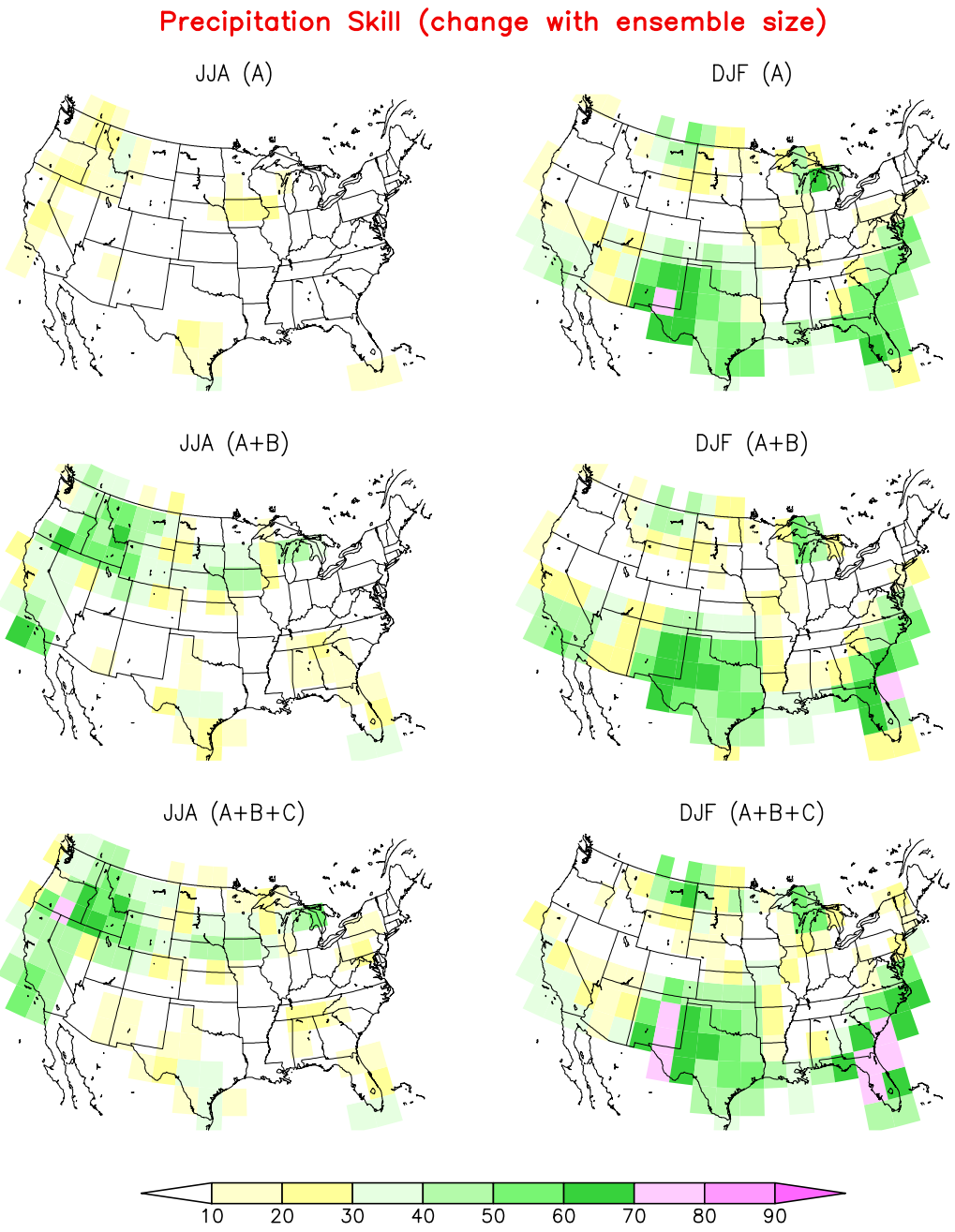

Fig. 11 As Fig.10, but now for precipitation.

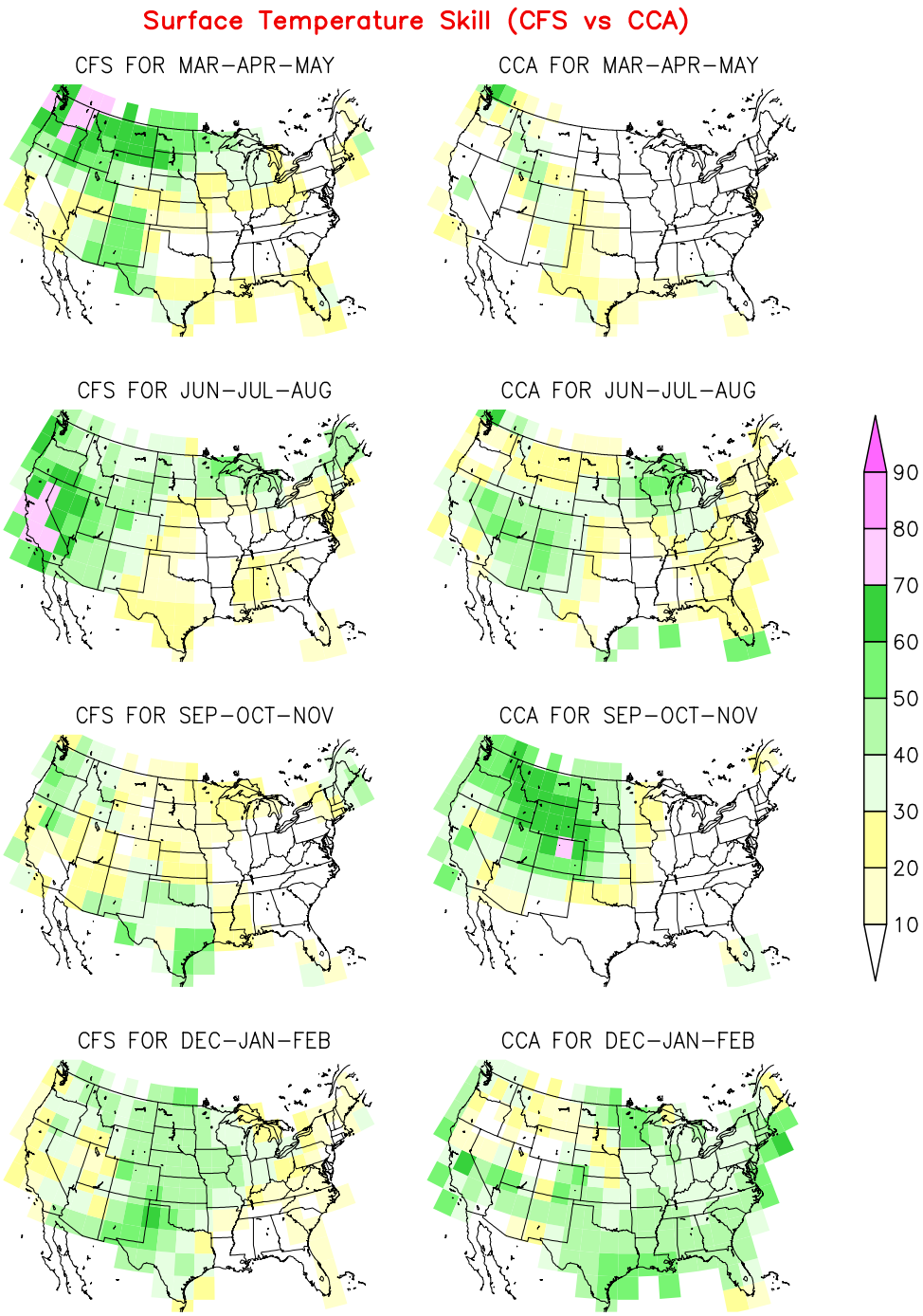

Fig.12 Left panels : Spatial distribution of retrospective ensemble mean CFS forecast skill (anomaly correlation in %) for lead 1 seasonal mean 2-meter temperature over the continental United States. The target seasons are, from top to bottom, MAM, JJA, SON and DJF. Right panels are the same, but for CCA. Note that CCA is based on a longer period, 1948-2003. Correlations less than 0.3 are in faint yellow and white.

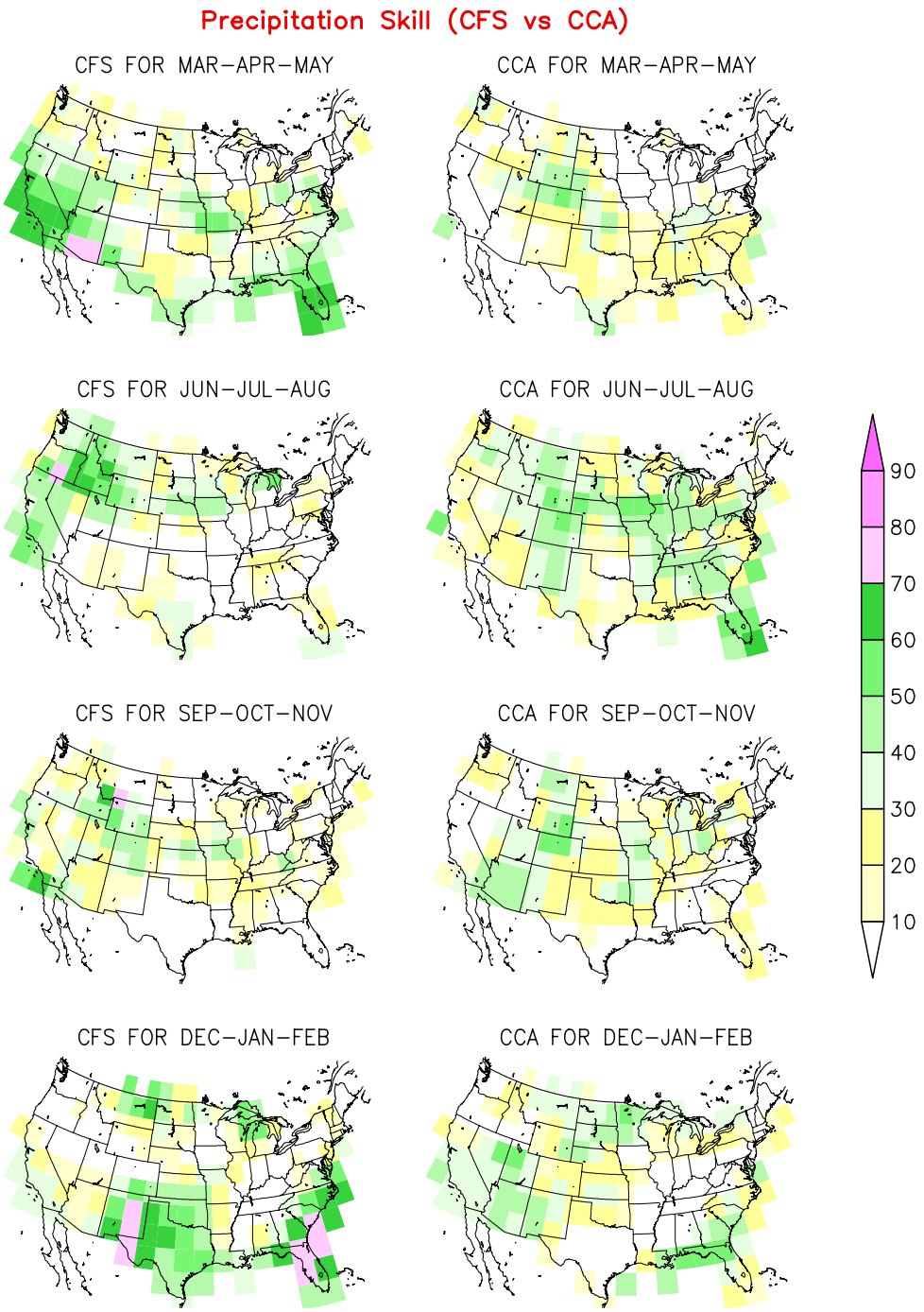

Fig. 13 The same as Fig.12, but now precipitation.

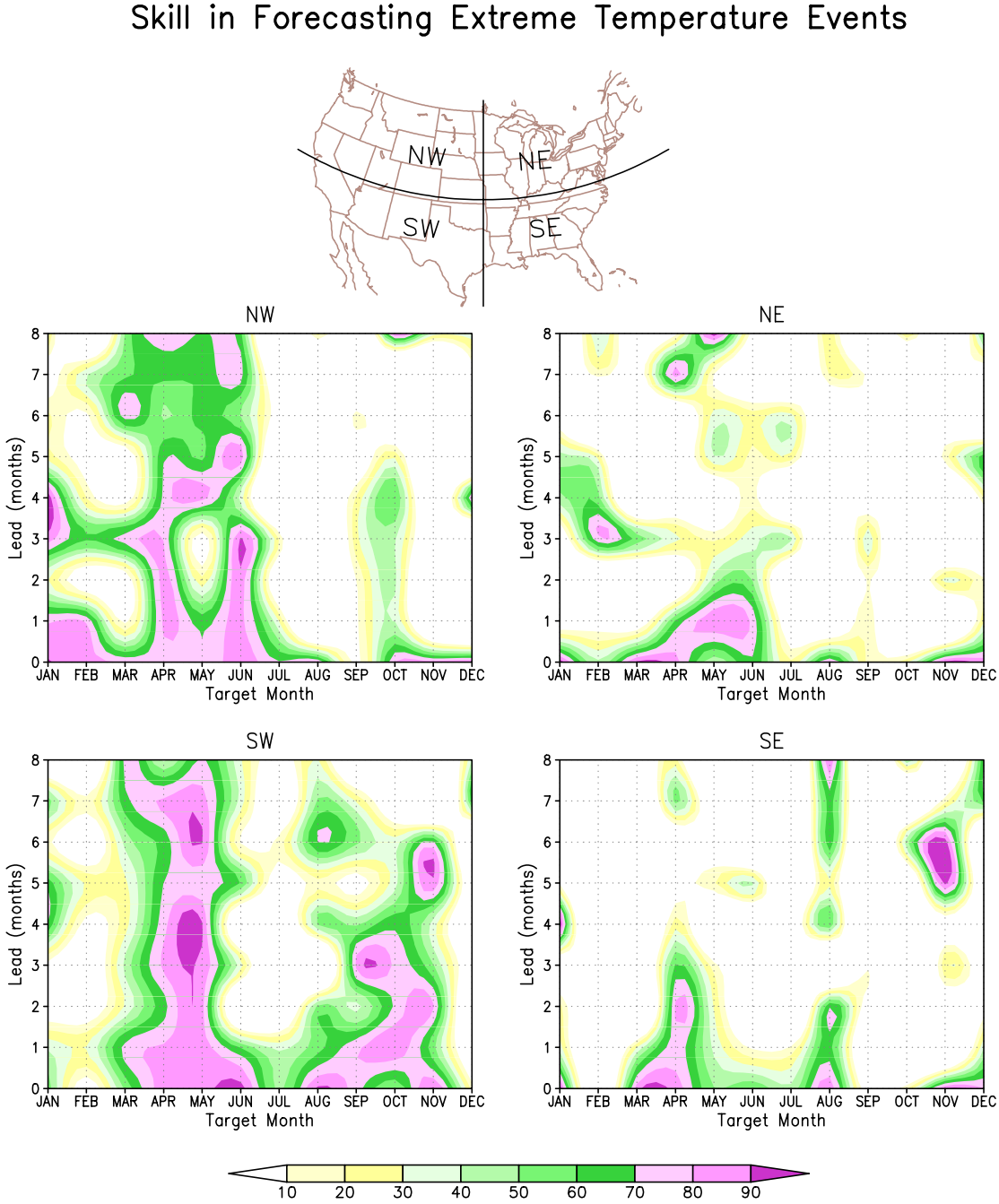

Fig. 14 Anomaly correlation (in %) of ensemble mean CFS forecasts as a function of lead (vertical) and target month (horizontal) for monthly mean 2-meter temperature over four quadrants of the continental United States (using 95°W and 37.5°N to define quadrants, see map at the top), evaluated only over those instances during 1981-2003 when an anomaly larger than 2 standard deviation occurred in the observations (anywhere in the quadrant). The much reduced sample size (relative to Fig.7), causes noisier patterns.

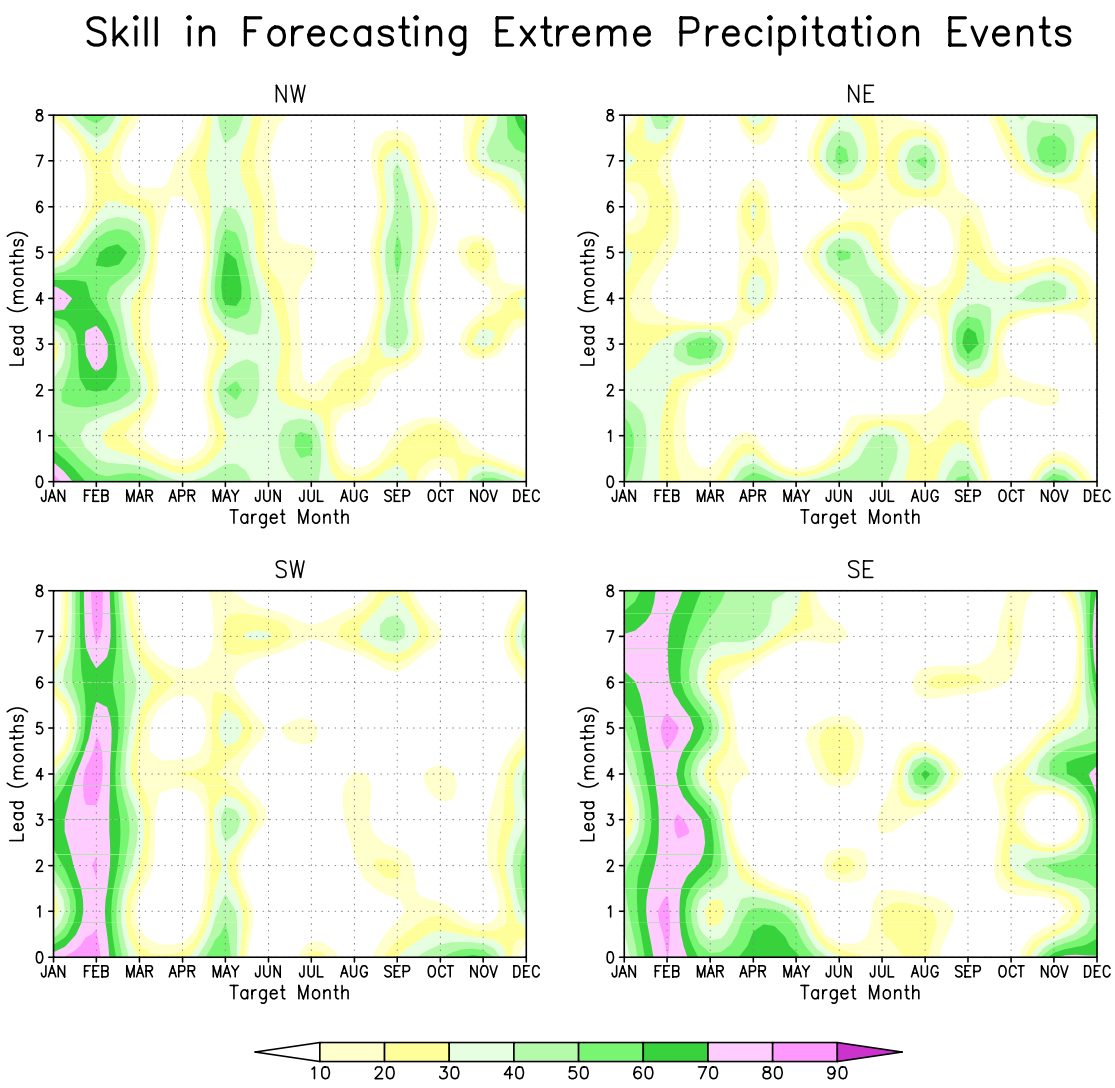

Fig. 15 The same as Fig. 14, but now precipitation.

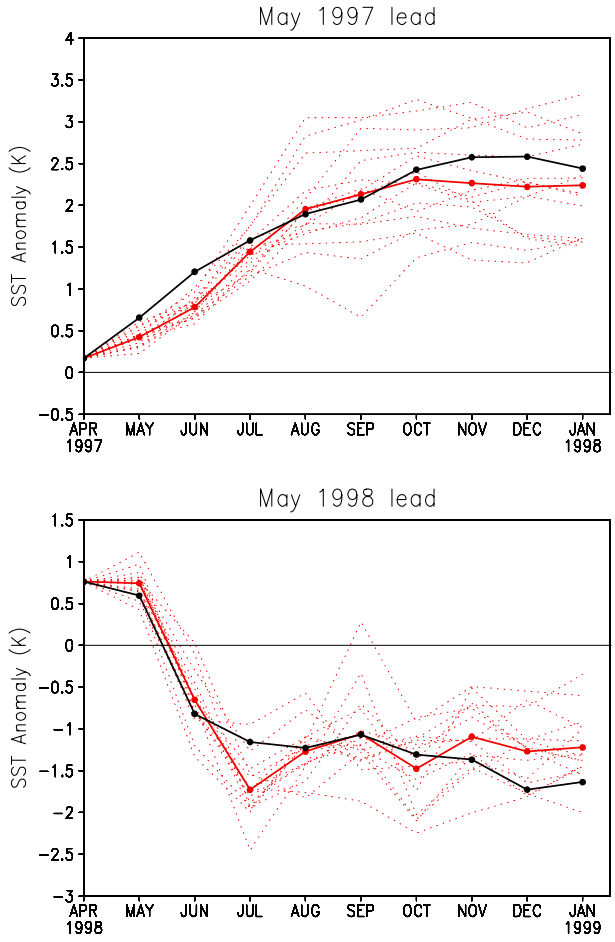

Fig. 16 Forecast plumes of Nino3.4 SST anomalies (^{o}K) from 15 initial conditions (from April 9 to May 3) in 1997 (top) and 1998 (bottom). All members are shown by red dotted lines, the ensemble mean is a full red line, and the observations are shown as a full black line.

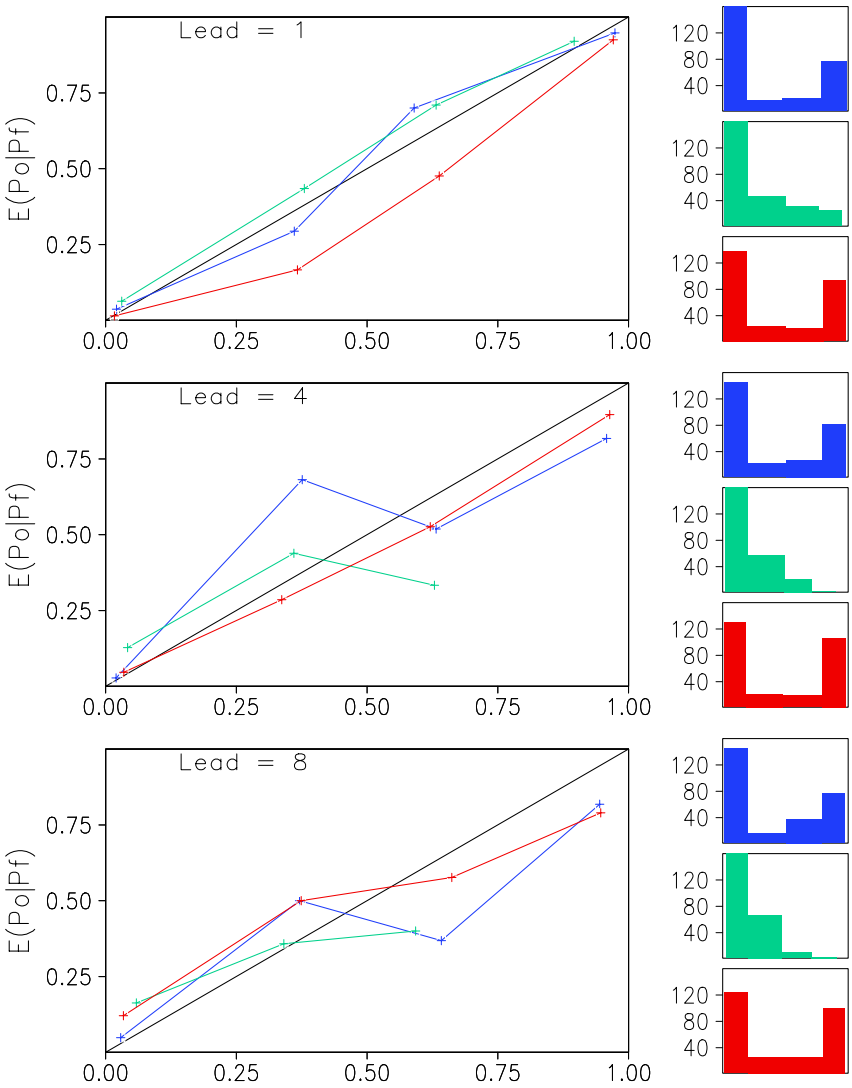

Fig. 17 Reliability diagrams of CFS forecast probabilities that Niño 3.4 SST predictions falls in the upper (red), the middle (green) and lower (blue) terciles of the observed climatology for (top) lead 1, (middle) lead 4 and (bottom) lead 8 months. The histograms on the right indicate the frequency of forecasts with probabilities in the ranges 0.0-0.25, 0.25-0.50, 0.50-0.75 and 0.75-1.0. Red colors correspond to forecasts for the upper (warm), green to the middle (neutral) and blue to the lower (cold) terciles. The black line (perfect reliability) is for reference.

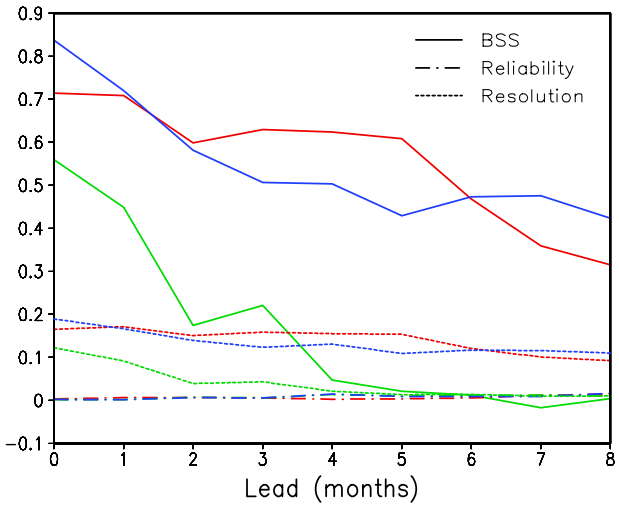

Fig. 18 Brier skill score (BSS), full lines, Reliability (dash-dot) and Resolution (dashed) as a function of lead time for three events: that SST in Nino3.4 is in the above tercile (red), in the middle tercile (green) and lower tercile (blue curves).

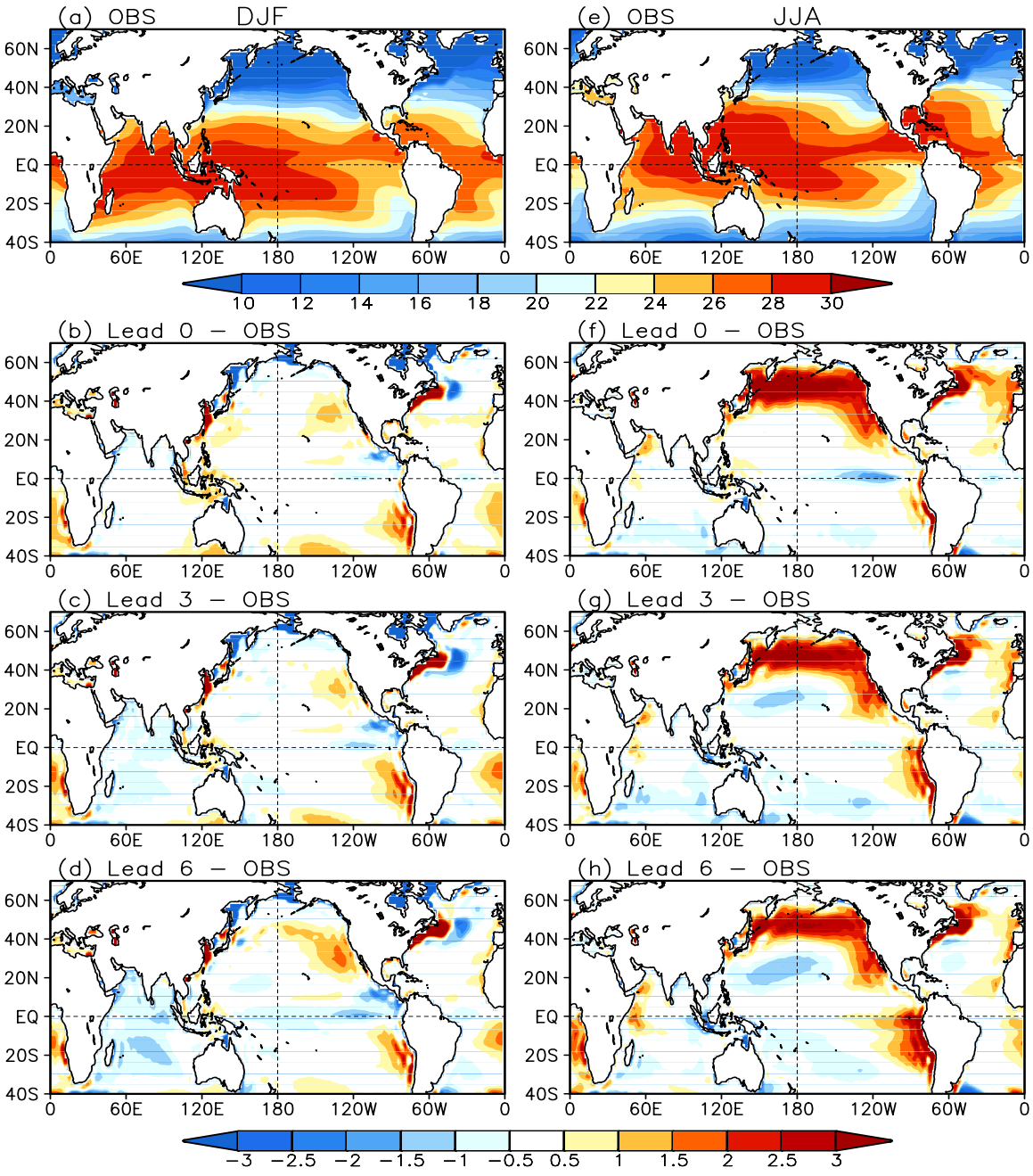

Fig.19. Observed climatology and the CFS model climate drift for SST. The climatology is defined over the period of 1982-2004. The climate drift is obtained by subtracting the observed climatology from the model forecast climatology. Left panels are for the winter season (DJF) and right panels are for the summer season (JJA). The top panels (a) and (e) are the observed climatology. The lower panels are the model climate drift for the 0-month lead, the 3-month lead and the 6-month lead, respectively. Unit is ^{o}C.

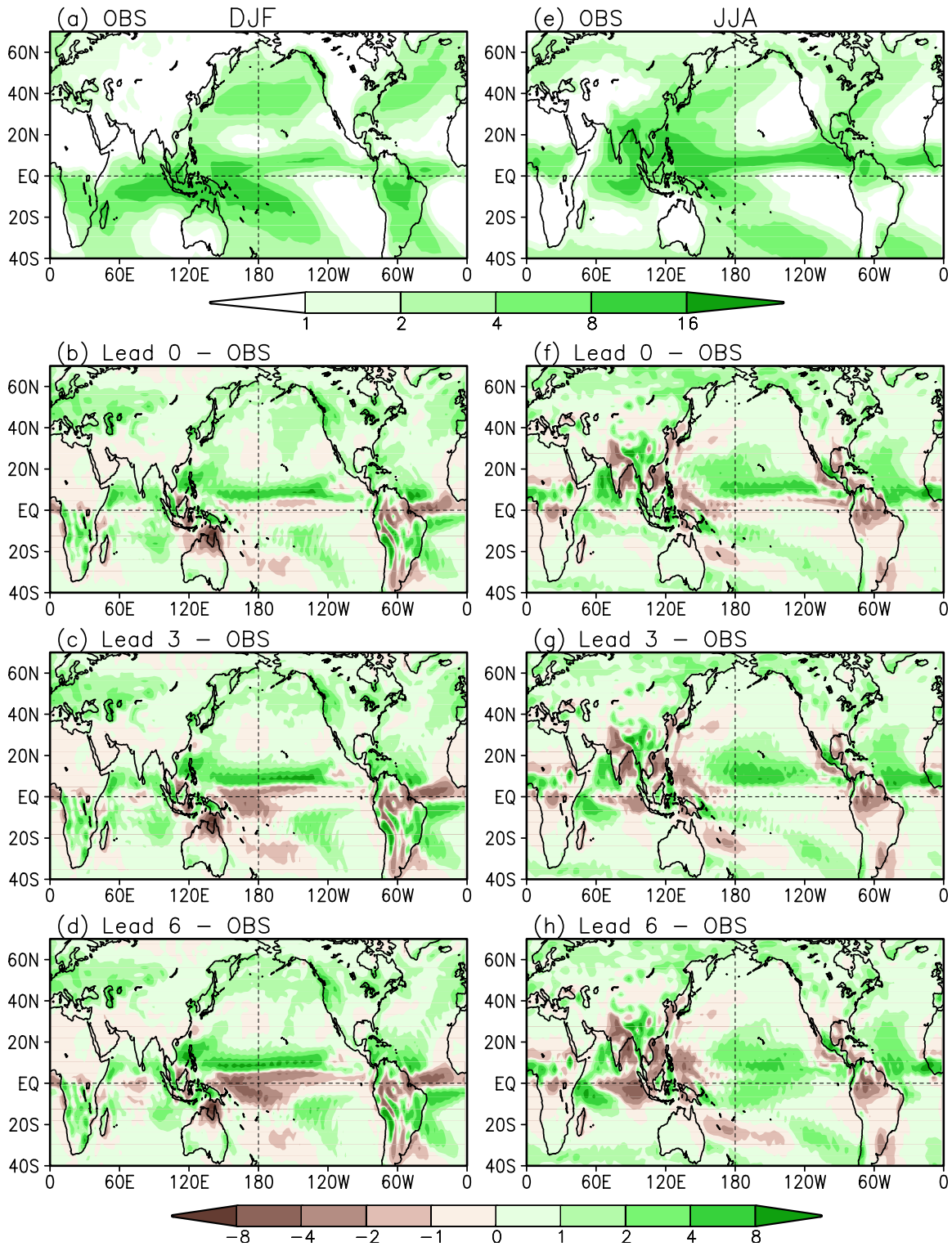

Fig.20. As in Fig.19 but for precipitation rate. Unit is mm/day

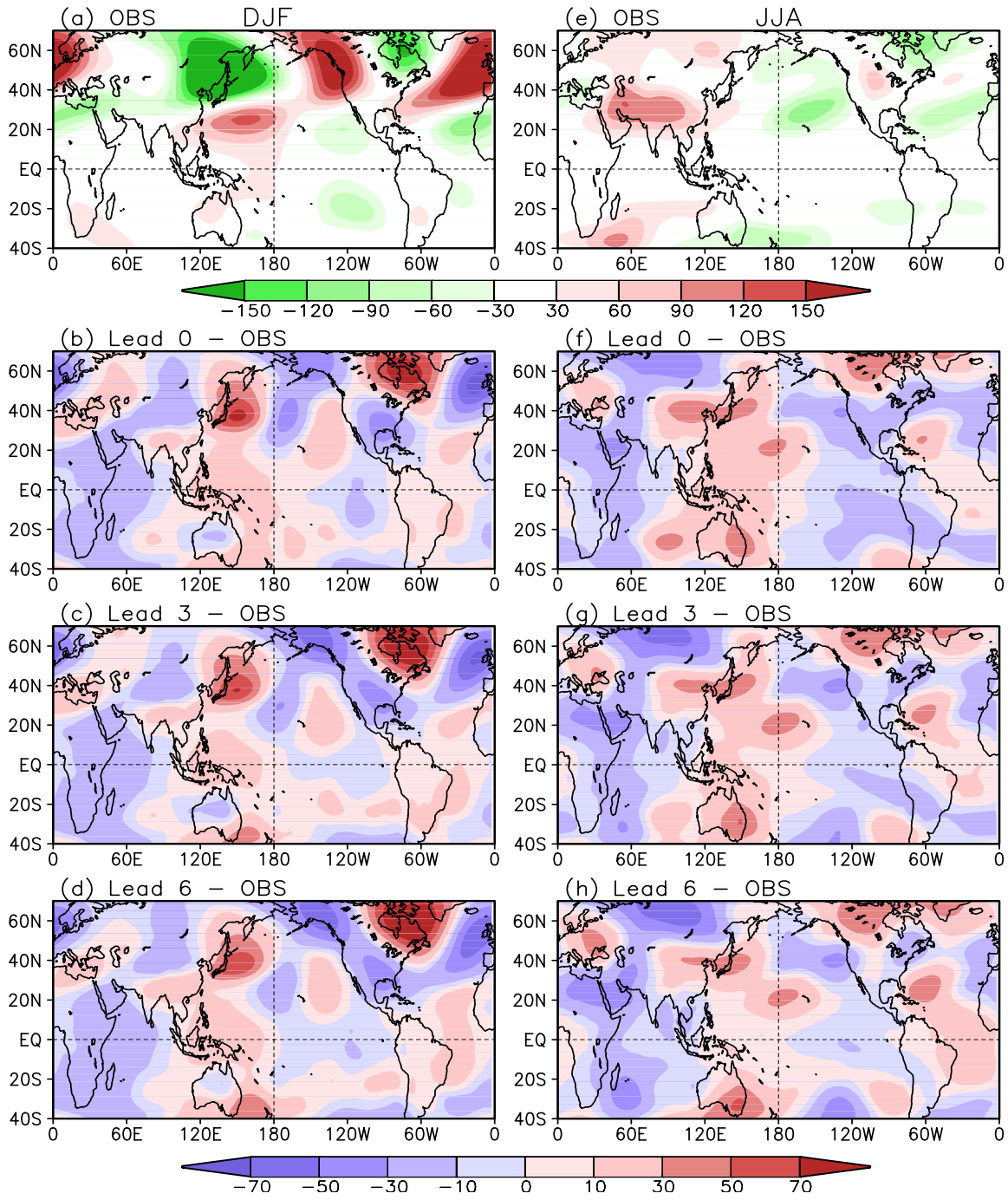

Fig. 21. As in Fig.19 but for 200hPa eddy geopotential. Unit is meters.

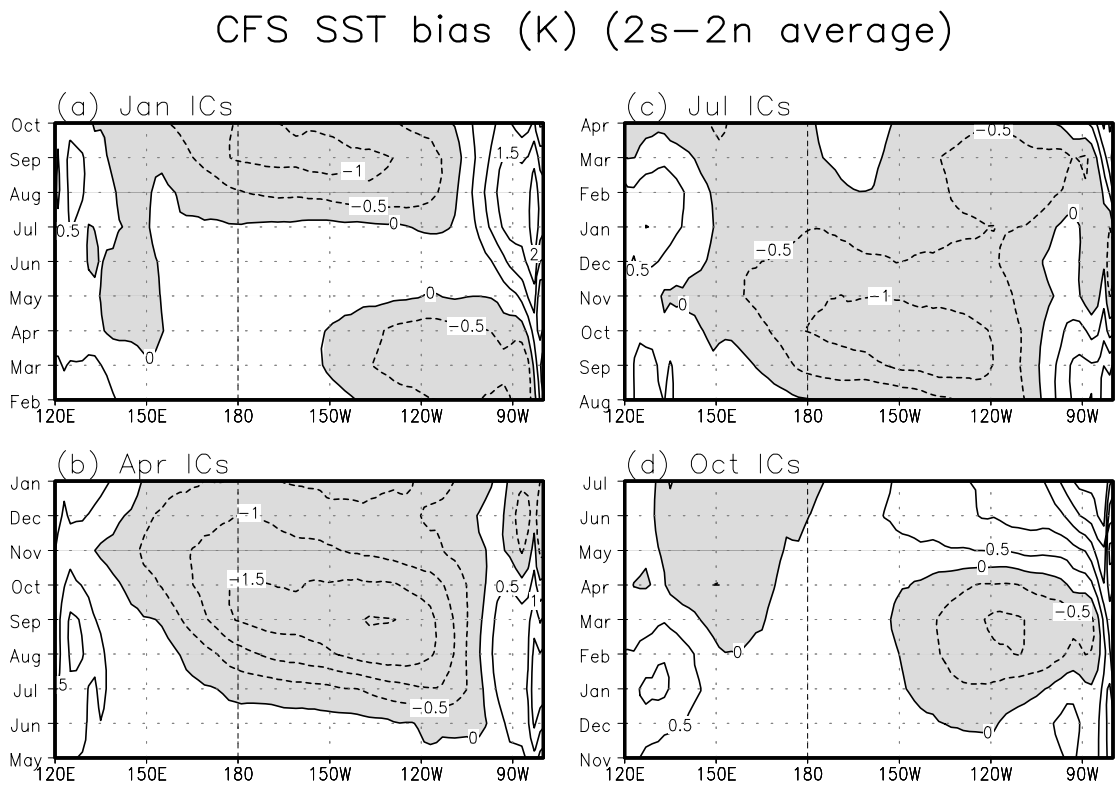

Fig. 22. Climate drift (Bias) of 2°S-2°N average SSTs in the Pacific for forecast from initial conditions of (a) January, (b) April, (c) July, and (d) October. Contours are drawn at 0.5 K interval. Negative values are shaded. The SST bias is relative to monthly OIv2 fields averaged over 1982-2004.

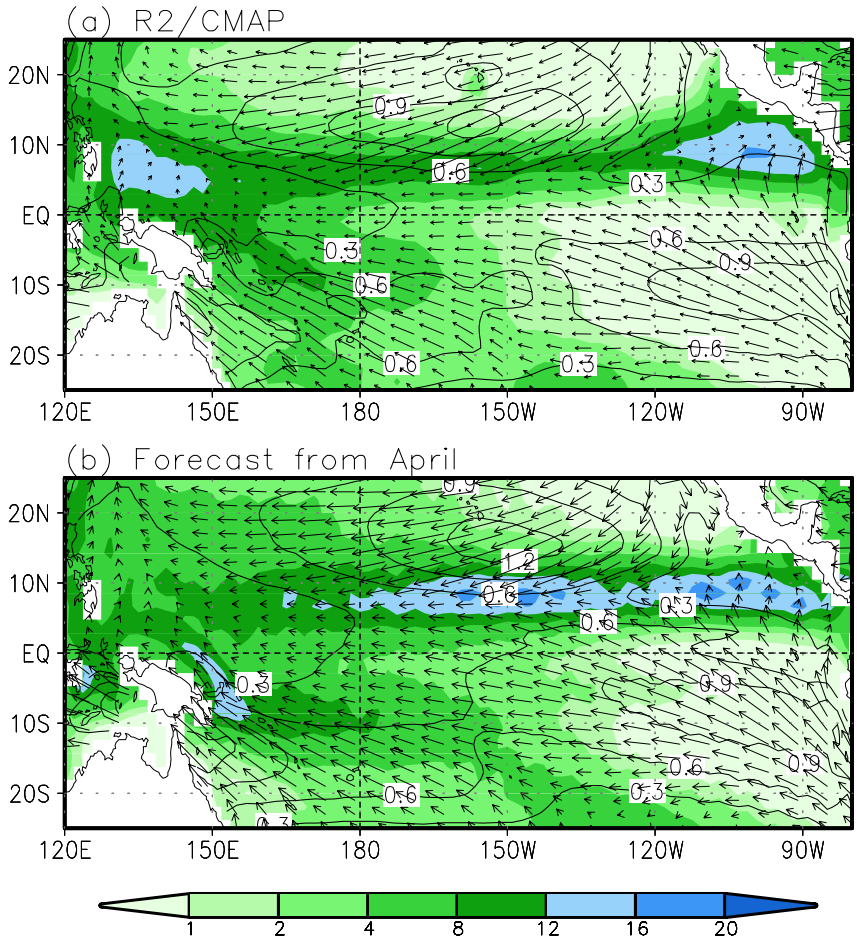

Fig. 23. Precipitation rate (color shadings) and surface momentum flux (vectors) for June from (a) R2/CMAP, and (b) CFS forecast from April initial condition. Contours are the amplitude of surface momentum flux (0.1 N m^{-2}). Precipitation rate is shaded at 1, 2, 4, 8, 16, and 20 mm day^{-1}.

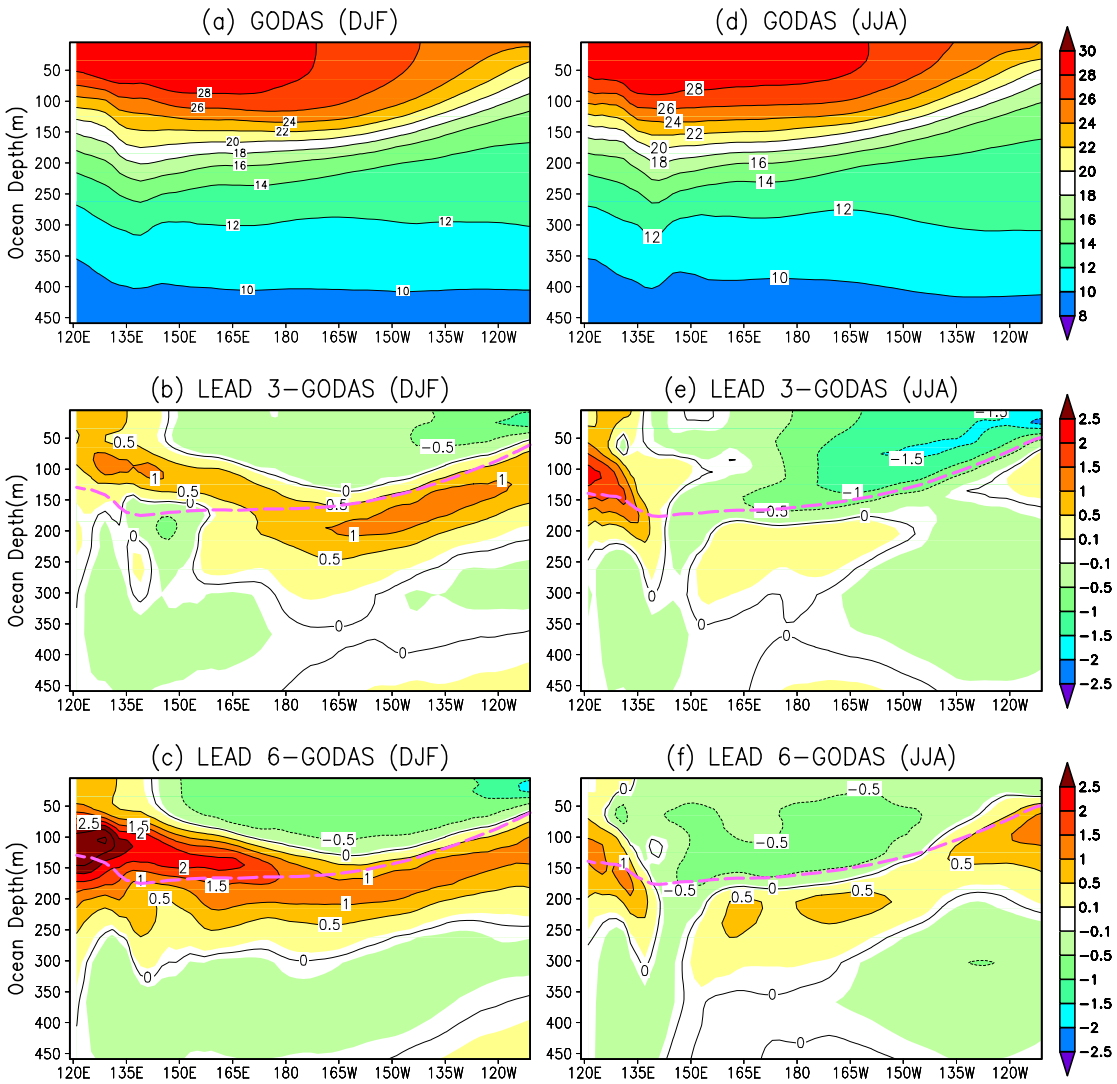

Fig. 24: The climatology of GODAS subsurface temperature in a depth-longitude cross section along the Equator in the Pacific and mean difference between the forecasts and GODAS in degrees Celsius. The figures on the left are for boreal winter (DJF), while the figures on the right are for boreal summer (JJA). The top panels show the climatology of subsurface temperature from GODAS. Note that a different scale is used for the color bar in the top panel. In the middle and bottom panels, the dashed pink line marks the 20C isotherm.

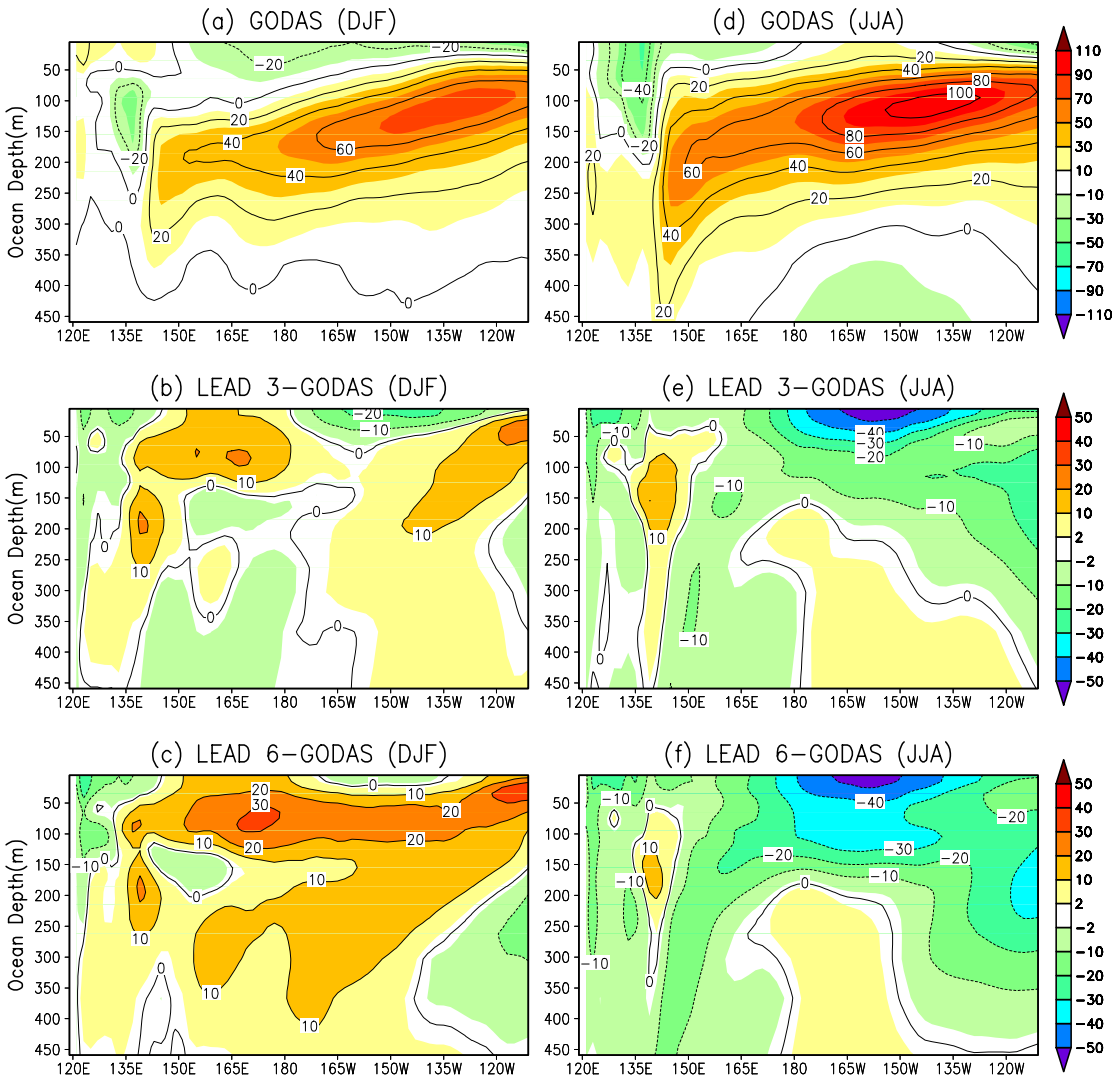

Fig. 25: As Fig. 24 but now zonal velocity in cm/s.

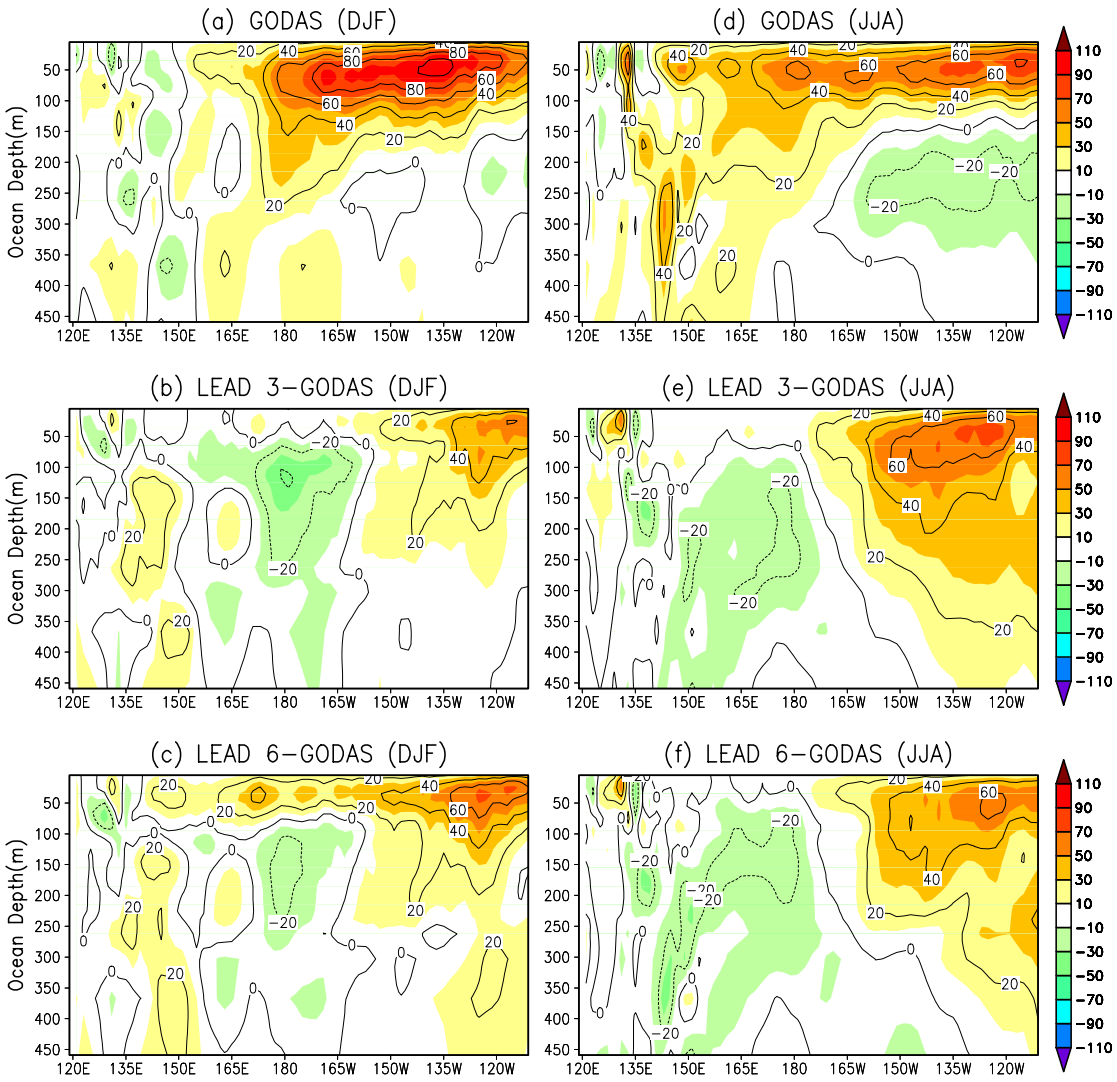

Fig. 26: As Fig. 24, but now vertical velocity in mm/hour

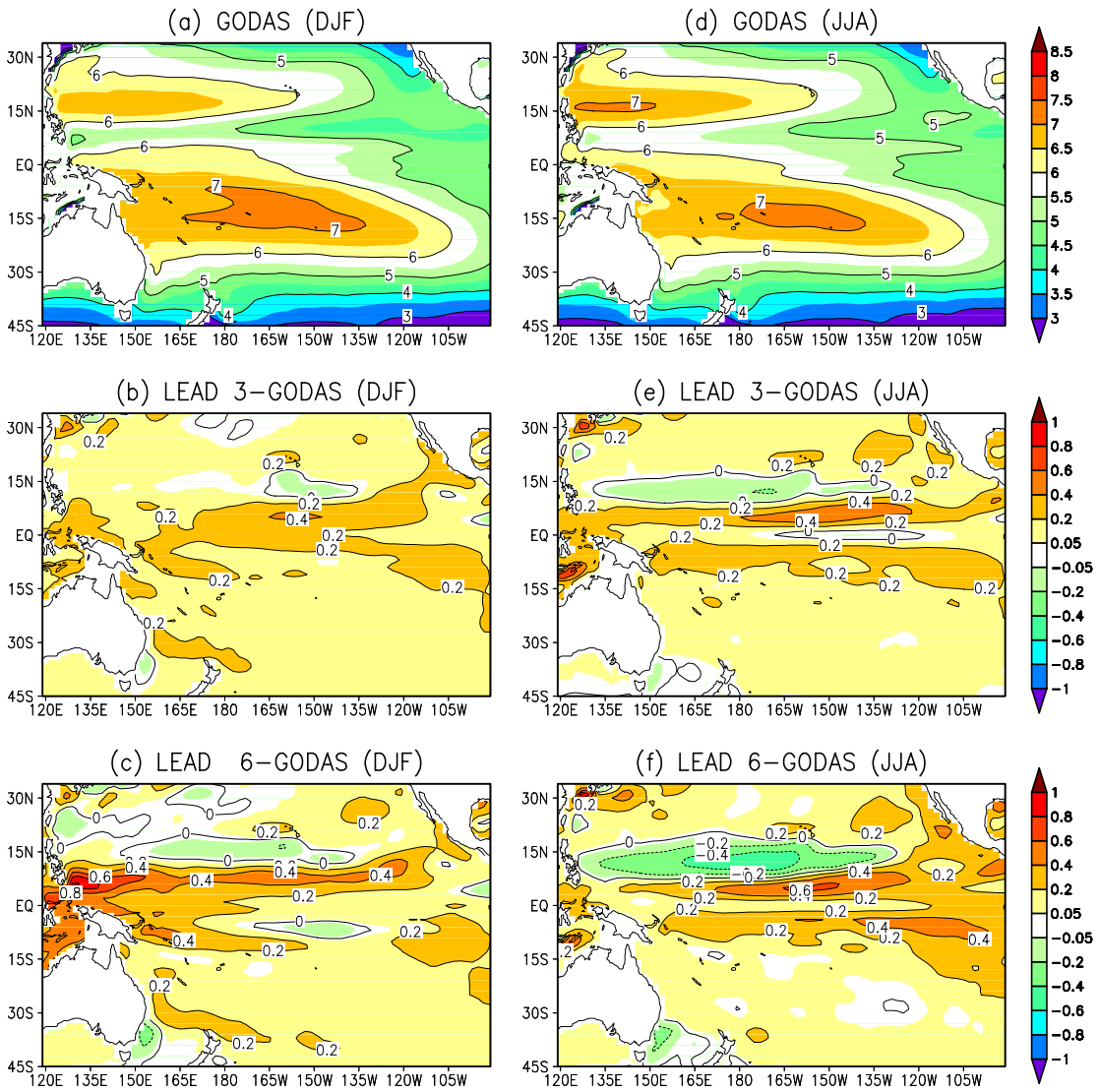

Fig. 27: As Fig. 24, but now a latitude/longitude representation of the upper ocean heat content in 10^7 J m^{-2}.

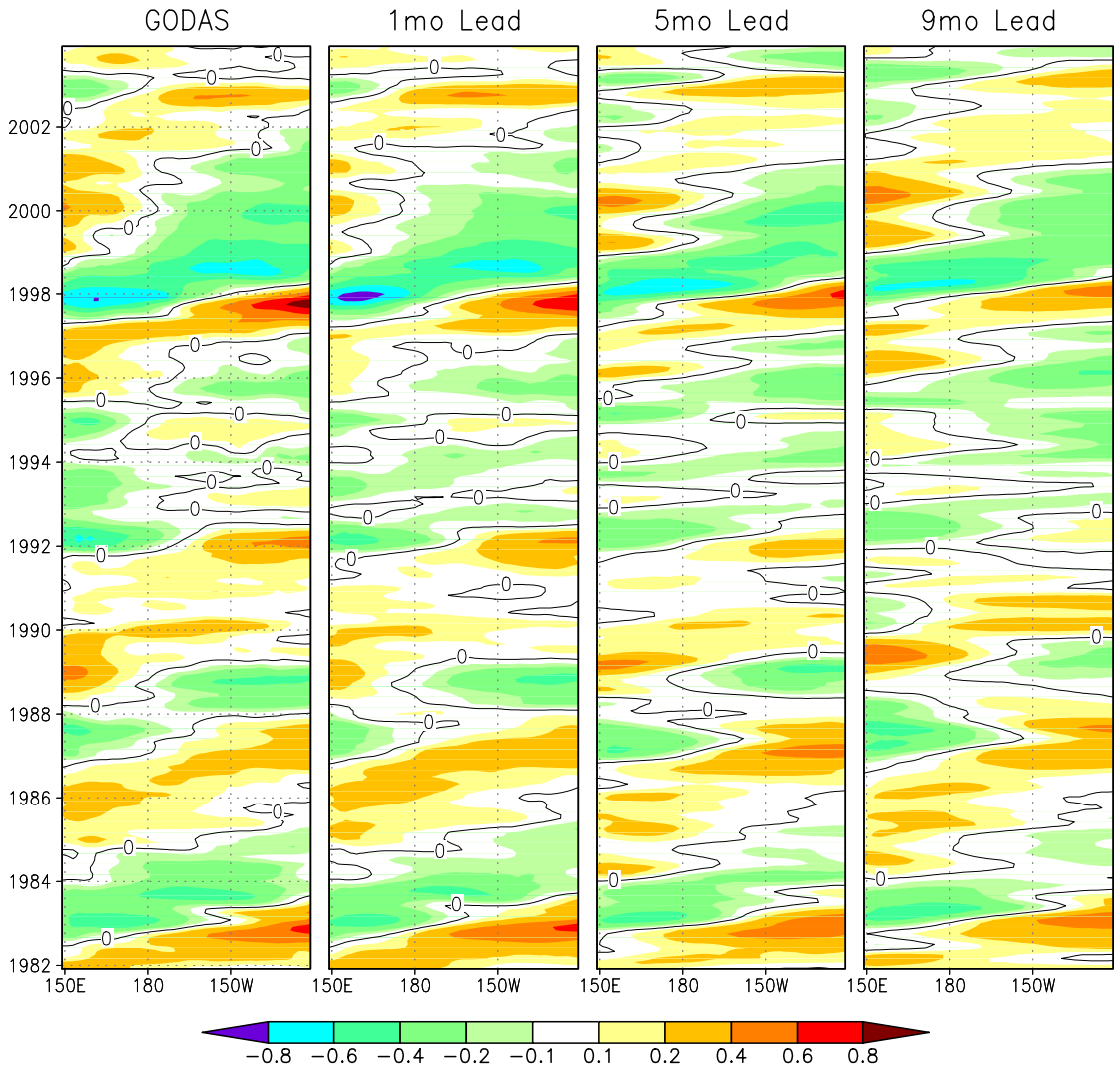

Fig. 28 Longitude-time plots of heat content anomalies along the equator in the Pacific from GODAS and CFS retrospective predictions. The climatology was computed for the period: 1982-2003. Unit is 10^7 J m^{-2}.

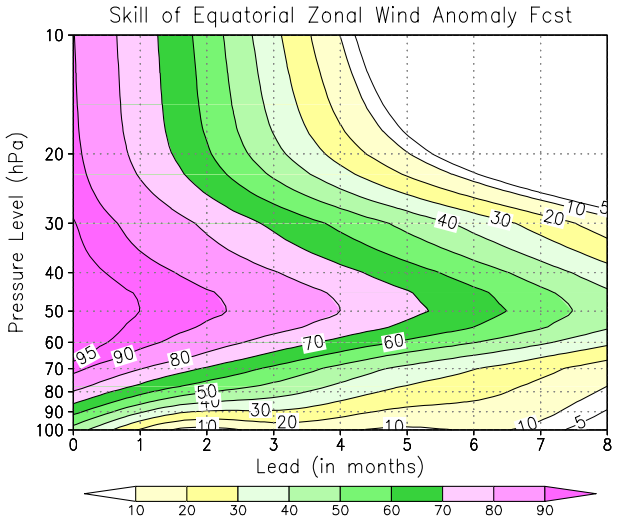

Fig. 29: Anomaly correlation (%) of Zonal mean zonal wind anomaly at the equator as a function of pressure level (above 100 hPa) versus forecast lead time (in month).